METHODS METADATA: CURATING SCIENTIFIC RESEARCH DATA FOR REUSE

BY

TIFFANY C. CHAO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Carole L. Palmer, Chair, University of Washington
Associate Professor Catherine L. Blake
Professor Jane Greenberg, Drexel University
Professor Michelle M. Wander

ABSTRACT

The sharing and reuse of research data relies on the provision of metadata and documentation. At present, scientists are not prepared to invest in the kind of rich description that will offer high functionality and interoperability in large-scale networked data systems. The absence of metadata not only makes data difficult to share and use but also increases the likelihood that these data will be lost, thereby diminishing the validity of the research data are associated with and leaving the potential value of the data for reuse unfulfilled and wasted. Further exacerbating this problem is the growing amount and heterogeneity of research data being generated and increased expectations from the federal government and publishers for public access to data.

Current efforts of data curators and managers to procure metadata from these data producers are highly time-intensive and an alternative approach is needed. This study investigates how research practices for data production conducted in the Earth Sciences can provide information about data that can be used for metadata description. The research methods implemented by scientists are an essential part of the data generation process and are important for metadata inclusion. Using a case study approach, I address what metadata for methods, or *methods metadata*, can be derived from two different sources of evidence: qualitative interviews conducted with data producers and content analysis of journal articles collected across three subdisciplines of Earth Science. Research areas such as the Earth Sciences already have scientists producing and working with a variety of data in need of curation support. Having a better understanding of how description for data can be systematically generated from both direct engagement with scientists and from use of the unobtrusive approach of analyzing research publications provides insight into ways to improve and support metadata development for data curators in research library, data repository, and archive environments and enhance data for sharing and reuse.

ACKNOWLEDGEMENTS

Portions of this dissertation reflect earlier research carried out with Dr. Carole Palmer and Dr. Melissa Cragin and supported by a NSF DataNet award for the *Data Conservancy: A Digital Research and Curation Virtual Organization* (OCI-0830976), led by Principal Investigator, Sayeed Choudhury, Sheridan Libraries, Johns Hopkins University. At the University of Illinois, Data Conservancy efforts were led by Co-Principal Investigator, Carole L. Palmer, Center for Informatics Research in Science and Scholarship (CIRSS). That work also benefitted from discussions with other Data Conservancy Data Practices team members, including Nicholas Weber, Andrea Thomer, and Karen Baker.

This journey would not have been possible without the guidance of my committee members, encouragement from friends, and support from my family. Thank you to the members of my dissertation committee: Cathy Blake, for her early support of this work and whose insights I have greatly valued her over the years, Jane Greenberg, for her unwavering enthusiasm of this project and willingness to speak at all hours, and Michelle Wander, for helping me make sense of the 'science' in this study. I am grateful to each of you for investing your time and expertise in this project and appreciate your thoughtful questions that have helped to advance this research. My deepest gratitude goes to my advisor, Carole Palmer, for her exceptional mentoring, tireless patience and caring, and for providing me with an intellectual home to grow as a scholar and researcher. Her commitment to the highest standards of research and teaching inspired and motivated me. I continue to strive to live up to her high standards as I embark on my professional career.

CIRSS has been a second home for me providing rich opportunities to participate in cutting-edge research and engage with world-class scholars. It is a place like none other and I feel fortunate to have been surrounded by so many talented and generous colleagues. I also thank the GSLIS front office staff and Help Desk/ITD for always responding to my inquiries, no matter how trivial they seem.

I am eternally grateful to my friends for ensuring that I had a place to write, reviewing drafts and offering feedback, and making sure time for frozen yogurt and custard breaks was a regular part of my academic experience. Being in different time zones was never a barrier and I cannot thank them enough for being a sounding board even when my ideas were half-baked. Sincerest thanks to my family for always encouraging me with their best wishes in my endeavors. A special thank you to my grandmother, Liang Chien-Chih, for her optimism and

positive outlook as I ventured into the research realm. I am very lucky to have the support of such caring people.

Finally, I acknowledge the institutional support that I have received while working on this research. In particular, I thank the Association for Information Science & Technology for the Thomson Reuters Doctoral Dissertation Proposal Scholarship, Graduate School of Library and Information Science, and CIRSS for supporting me with generous fellowships and making dissertation writing and conference travel possible.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The need to support long-term access and reuse of research data has made data curation a priority for research libraries, data centers, and repositories. A key part of the curation process is ensuring that metadata is available to describe datasets in order to facilitate future use. However, the provision of metadata by the data producer, who best understands how and why data are gathered, is not always a common practice or cultural norm (Karasti, Baker, & Halkola, 2006). Funders have long called for their grantees to collect and maintain metadata, but this call has been met with minimal adherence or completely ignored. As explained by Edwards et al. (2007), this "metadata conundrum represents a classic mismatch of incentives: while of clear value to the larger community, metadata offers little to nothing to those tasked with producing it and may prove costly and time intensive to boot" (p. 32). Metadata for long tail-science research data are a particular concern since these data are quite heterogeneous and often, a single metadata standard cannot be readily applied. Examples of long-tail science include areas of study within the Earth Sciences, particularly with field-based work, where scientists often lack the financial support or tools for metadata generation thereby limiting future access and reuse of data produced (Heidorn, 2008).

Since data producers are not necessarily volunteering metadata, those providing data services often need to generate the information by other means. The "data interview" is one technique that is steadily becoming a part of research data services within academic libraries (Witt, Carlson, Brandt, & Cragin, 2009; Carlson, 2012). It is a productive approach for gathering information about practices and expectations surrounding data production and long-term access and other information needed for producing descriptive metadata. However, interviews are time consuming and can be challenging to arrange and conduct with busy researchers. More efficient means for generating metadata are needed to streamline curation processes and provide adequate information for data to be reused by others.

The use of semi-automated approaches to generate metadata can be an effective technique that yields an initial foundation of descriptive metadata for research data without creating a large burden on data producers. Automated approaches for harvesting and extracting bibliographic metadata have been applied to text (e.g. Kovacevic et al., 2011) yet necessitate a degree of human intervention to ensure quality documentation (Greenberg, 2004). For research data, the journal publications of data producers are one of the dominant

modes for communicating scholarly information within scientific communities and could be a rich source of content for generating metadata for datasets. Scientific journal publications remain a primary mechanism of communication amongst scientists and scholars for disseminating scientific knowledge (Brown, 2010). Moreover, with the increase of the number of open access journals, published articles are more readily available than in the past. The representations of data (e.g., figures, tables, charts, etc.) and narrative content embedded in journal articles, particularly descriptions of methods implemented in the research, plays a vital role for researchers in validating the reliability of data for reuse (Faniel & Jacobsen, 2010). Data underlie the results published in journals, and they are increasingly made accessible as supplements to published articles (Borgman, 2012) or deposited in domain repositories in response to journal publisher policies, further emphasizing not only the role of articles in representing the data but also the significance of linking a research publication with its respective dataset. Overall, the contents of journal articles present a significant information source for use in generating descriptive information for research data.

Establishing "methods metadata"

Utilizing journal publications for metadata production requires an understanding of the practices related to data production and what aspects of these practices need to be represented in the metadata describing a dataset. As stated by Gray, Szalay, Thakar, and Stoughton (2002) "(d)ata is incomprehensible and hence useless unless there is a detailed and clear description of how and when it was gathered, and how the derived data was produced" (n.p.). For this dissertation, I introduce the term *methods metadata* to signify the type of information needed for basic comprehension of how data were produced and analyzed in the scientific research context. This metadata includes descriptive information on any procedures undertaken in the gathering, collecting, processing, or analysis of research data.

Formal definitions of "methods" provided an initial basis in conceptualizing this term for descriptive metadata purposes. In its singular form, "method" is defined in the OED as "a special form of procedure or characteristic set of procedures employed (more or less systematically) in an intellectual discipline or field of study as a mode of investigation and inquiry, or of teaching and exposition."[1] The use of the plural "methods" is intentional for this

[1] Method [Def. 3]. (n.d.). In *OED Online.* Retrieved from http://www.oed.com/view/Entry/117560?rskey=TLRwdw&result=1&isAdvanced=false#eid
[2] Data Conservancy project site: http://dataconservancy.org/

study to denote the multiple procedures and techniques that drive the different stages of data production and analysis. To understand the role of methods in the context of scientific research data, we can turn to methods-related metadata elements from Earth Science metadata schemes. The explanation for "Method" from the PANGAEA repository-based metadata does not include any discussion of procedure, instead asking for the full name of the method and associated publication describing it. The "Method-specific metadata" from the EarthChem repository-based metadata also lacked a formal definition for "method(s)" but rather, details specific information to be included in describing a method for analysis of samples (i.e. detection limit). In contrast, the "Methods" definition from the Ecological Metadata Language is stated as "the *actual* procedures that are used in the creation or the subsequent processing of a dataset." With the emphasis on "procedures" from these formal definitions, this research further contributes the importance of underlying research practices in data production and analysis to more fully elucidate the meaning of methods metadata for research data.

Research questions and approach

The primary aim of this research is to inform metadata generation processes for data curation services. A qualitative approach was applied to examine the data production practices from three Earth Science subdisciplines representative of long tail science research: Soil Ecology, Volcanology, and Stratigraphy. Two sources of evidence were used in the study: 1) semi-structured interviews covering practices of data production in the three subdisciplines, and 2) content analysis of journal articles published in these subdisciplines.

The qualitative interviews are based on preliminary work on data practices in the Earth Sciences I conducted as a member of the Data Practices group, a research team within the Data Conservancy[2] initiative. The Data Conservancy as a whole performed research and developed data infrastructure and services to foster preservation, sharing, and discovery of research data across institutions and disciplines. The Data Practices team carried out qualitative studies of Earth and life sciences to support Data Conservancy efforts. The varied and complex data types and formats produced in Earth Science research, in particular, were ideal for examining the breadth of data types and practices with implications for curation services. One of the primary outcomes of the Data Practices group was the development of a vocabulary to describe the

[2] Data Conservancy project site: http://dataconservancy.org/

relationships between curation activities, data, and research practice (Chao, Cragin, & Palmer, 2015). The group also focused on the reuse potential of data across subdisciplines and multiple functions of different data type groupings (Palmer, Weber, & Cragin, 2011; Palmer et al., 2012).

For this dissertation, I conducted an original analysis of the qualitative data collected in interviews with Earth Sciences researchers from the Data Practices research by addressing new research questions and adding a second phase of research through the content analysis of published journal research articles. This dual technique differs significantly from the approach used by the Data Practices team, particularly in the emphasis on descriptive metadata and research methodologies represented journal articles as a source of information for describing data. I investigated the following set of interrelated questions:

A) What methods metadata relating to how data are gathered, processed, and analyzed for research can be derived from:
    - qualitative interviews?
    - research journal articles?

B) How does methods metadata differ across subdisciplines in Earth Science?

The specifics of how data are gathered, processed, and analyzed, identified in interviews and journal articles, provided insights on data production and analysis processes essential to methods metadata, and on identifying variations across the three Earth Science subdisciplines. As a complement to the qualitative interviews with Earth Science researchers, I used journal articles published by these researchers along with a set of articles gathered on cognate research for each subdiscipline in the investigation of methods metadata. The analysis of the interviews and articles in this first phase of research resulted in the formation of a set of terms specific to describing methods (Methods Metadata Categories). I also examined how the interviews and articles contributed to the identification of methods metadata for each subdiscipline. The set of terms was corroborated in the second phase of research with multiple metadata schemes for scientific data in the Earth Sciences to account for the range of subdisciplines examined. The content of metadata records from data repositories applying these metadata schemes was also investigated to understand how methods descriptions have been represented. I describe subdiscipline observations for each phase of research, which are brought together and discussed in the third, and final phase of research.

Contributions of research

Intellectual contribution

This research brings attention to the significance of methods description for data reuse and addresses how scientific practices, in particular data production and analysis processes, contribute to metadata development. The identification and description of methods metadata is a starting point in addressing the issue of mismatched incentives introduced by Edwards et al. (2007) and overcoming the friction of metadata generation experienced by scientists through recognition of the practices for gathering, processing, and analyzing data in scientific research. With scientists using different methods, this study can contribute an understanding of what information components are crucial for methods metadata description for research datasets.

Past studies of data practices across multiple disciplines (i.e. Swan & Brown, 2008; Lyon, Rusbridge, Neilson, & Whyte, 2010) have conducted analysis within broad disciplinary groupings that can mask important distinctions within research or subject areas. This is especially problematic for the long tail sciences where research processes and techniques used tend to be localized and unique to the small research team (Wallis, Rolando, & Borgman, 2013). By investigating methods and data production practices at the subdiscipline level, the potential range of similarities and differences were revealed to not only enhance how metadata schemes are used but also how they can be improved to better support access and reuse of research data across disciplines (Willis, Greenberg, & White, 2012). Methodologically, my study also furthers knowledge on the benefits and limitations of conducting analysis on interviews and journal articles at the subdisciplinary level.

Practical applications

A more effective approach for generating metadata for research data is important in developing tools for facilitating data repository deposits. The combined technique of direct interviewing of scientists in addition to the unobtrusive approach of identifying metadata through journal articles addressed in this study could potentially be adopted by data curation professionals to secure description about methods for other research areas. The current approach of data interviews can yield thorough results for information about how data were generating but can be limited based on available time and response from data producers. The examination of methods description in journal articles therefore lends a new perspective to

more structured and even automated approaches for metadata extraction of datasets (Greenberg, White, Carrier, & Scherle, 2009). This is especially valuable for enhancing metadata generation for long tail science research fields that do not regularly use instrumentation that automatically documents the processes undertaken when collecting and processing data (to contrast see Goble et al., 2008 for automated documentation in bioinformatics).

This research is particularly timely as demand for curation services is expected to escalate due to new expectations for public access to digital data produced by federally funded research, as outlined in the February 2013 Office of Science and Technology Policy memo (Holdren, 2013). Research techniques such as meta-analysis are also gaining traction in the long tail science research areas, including Earth Science (i.e. Koricheva, Gurevitch, & Mengersen, 2013), which increases the need for discovery and access of data from multiple studies with common parameters. Moreover, metadata about the research methods used to generate the data for each study will be critical for determining the applicability and compatibility of data for integration and re-analysis.

Structure of the dissertation

The next chapter situates this dissertation in the context of relevant literature on data practices in the long tail sciences with a focus on data reuse and sharing, metadata generation and standards, and differences in practices across multiple disciplines. Chapter 3 follows from the review of background literature to describe the overall research design for this dissertation to identify methods metadata for three Earth Science subdisciplines. In Chapters 4 and 5, I present the analysis results. Chapter 6 provides a summary of the primary findings from each phase of analysis in response to the research questions. This chapter concludes with an examination of the limitations of this study and directions for future research.

CHAPTER 2: BACKGROUND LITERATURE

The research data produced in the "long tail" of science have immense potential to foster scientific innovation. However, due to limited resources for long-term maintenance and care, these data are largely unavailable for reuse and subsequently forgotten (Heidorn, 2008). There is a range of literature on metadata standards development for data and cross-disciplinary studies on data sharing perceptions and expectations, but most of it does not extend to the actual practices and experiences of scientists producing data in their research and how metadata is generated for data in this context. For the purposes of this study, where the focus is on metadata generation from scientific research methods and approaches to identify scientists' data production methods from two different data sources (semi-structured interviews and journal publications), studies on the research data practices of scientists offer insights into the data production process within research communities.

In the following sections I position methods metadata within the context of data curation and discuss relevant literature as it relates to major themes represented in my research questions. Related work includes studies of metadata generation for data reuse in practice studies, disciplinary differences in research data practices, and studies research data sharing in the long tail sciences. To situate the first research question, I focus on scientists' data production processes, drawing on findings from research practice studies conducted in the long tail sciences and discussing the implications for metadata generation and the identification of research methods. Empirical studies on the data-oriented practices of long tail science researchers are still emerging within a richer body of literature developed on data management and curation. To position the second question on methods metadata variation across Earth Science subdisciplines, I continue to focus on practice studies specific to disciplinary differences and discuss implications for my research design. I conclude with a discussion of *metadata frictions* and *circulating reference* as two conceptual frameworks from the social studies of science literature to guide the design and analysis of this research.

Situating "methods metadata" in data curation

Description for data and datasets as a curatorial activity is associated with several interrelated terms within the curation literature: *documentation*, *metadata*, and *representation information*. The provision of *metadata* and *documentation* consistently appear as products of curation activities in models of the data lifecycle (Ball, 2012), and the terms seem to have

preferred use depending on discipline. For instance, *documentation* is favored in the social science data curation community, where the Data Documentation Initiative (DDI) metadata schema is an important standard for capturing the information necessary for data discovery and analysis of quantitative social science data (Vardigan, Heus, & Thomas, 2008). The DDI standard builds on the traditional codebook and document-centric forms of documentation in the social sciences and supports the representation of the lifecycle of data through metadata aimed at capturing transformations that occur throughout the research process. On the other hand, *metadata* is the more common term used in the natural science data management (Niu & Hedstrom, 2008).

The third term, *representation information* is part of the prominent Digital Curation Centre (DCC) Curation Lifecycle model. It is portrayed as "description and representation information" and defined as,

> "Assign[ing] administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect[ing] and assign[ing] representation information required to understand and render both the digital material and the associated metadata" (Higgins, 2008, p. 137).

Noticeably, the different types of *metadata* are included in the definition, and *metadata*, per se, does not appear as a separate element within the DCC Curation Lifecycle model. An enhanced understanding of *representation information* is detailed in the Open Archival Information System (OAIS) reference model: representation information is all encompassing information focused on providing meaningful context to a data object and elaborated through structural (how information is organized); semantic (further describes meaning); and other (i.e. software, algorithms, encryption) attributes (Consultative Committee for Space Data Systems [CCSDS], 2012). Both the DCC Curation Lifecycle model and OAIS model have gained considerable traction in the data curation community for the design and implementation of data systems and services.

Despite the variations in the names of terms, the common goal conveyed by these terms and definitions is the provision of sufficient information about the data to allow use and interpretation in a meaningful way by future users. The focus on "methods" as a specific type of metadata for inclusion makes visible the processes and transformations that data have undergone in the course of the research study. Emphasis on methods description also resonates with the related curation term *provenance*, which is described in OAIS. Within

scientific domains that are more computationally driven such as modeling, *provenance* and *lineage* are prevalent terms used to denote description of how data products were generated, including the scripts or programs used in the process (Bose & Frew, 2005). *Lineage*, in particular, not only accounts for the history of how a dataset originated but also what transformations or resulting products have emerged since the creation of the dataset. The specifications of these interrelated products and activities are akin to methods metadata though the terms are not as familiar within the long tail sciences. The significance of research methods and their importance for inclusion in metadata for scientific data is investigated and discussed in the following section.

Describing data for reuse

The role of 'methods' description in data reuse

In order to understand how methods metadata can be supported as part of data curation, it is necessary to examine how scientists conduct their research and produce data. Research practice studies related to the role of data provide a foundation of empirical evidence that helps us to recognize how and why certain activities are performed and decisions are made in the real world. Review of the limited practice studies specific to the long tail sciences not only reveals the critical role that methods information has in informing scientists of data quality, but also what information should be imparted through methods descriptions.

Description of methods and processes used to generate data provide scientists with insight into whether available data can be reused for new purposes. The research methods employed can convey the level of professionalism and expertise of the data producer within his or her scientific community (Faniel & Jacobsen, 2010). Researchers in the long tail science area of environmental sciences determine whether to trust the quality of environmental datasets by first evaluating the scientific processes that were employed in creating the data and then assessing the personal and professional reputation of the individual, group, or organization that produced the dataset in order to counteract any biases that the chosen methods for generating data may have (Van House, Butler & Schiff, 1998). Similarly, Zimmerman's (2008) study of ecological research practices uncovered that the documentation on methodologies was significant in appraising trust in data and guiding selection of data for reuse. The importance of information on methods and processes of data production is also a persistent theme in studies of data practices beyond the Earth Sciences. Methods and protocol information for genomics

research is often made available through project websites to complement a dataset deposited in GenBank[3], and assessment of methods deployed to produce data is common in the peer-review process for publications in astronomy research (Swan & Brown, 2008).

While the inclusion of research methods is an integral component for assessing data quality, it is also important to maintain descriptions of methods throughout the research process. In field-based sciences such as ecology, rapid environmental changes demand immediate decisions that may alter research methods and protocols in order to properly capture a particular phenomenon (Mayernik, Wallis, Pepe, & Borgman, 2008; Karasti, Baker & Halkola, 2006). Presenting an ecological research dataset for public consumption should include discussion of missing values, modifications during procedure implementation, or natural disturbances that occur in the ecological environment at the time of collection (Karasti & Baker, 2008). The absence of these documented changes may impact the overall integrity of the dataset.

Other attributes of value for inclusion in descriptions of methods relate to the research activities of data generation and analysis. Wallis (2012) depicts the variation in ecology between traditional practices of taking sensor readings by hand and the alternative use of networked embedded sensor technologies that automatically collect the same contextual variables. The adoption—or lack of adoption—of new technologies by a research team is an obvious difference in data gathering within a single research community that needs to be adequately documented in the metadata for a dataset. In other field-based sciences where physical samples are used for analysis, the inclusion of methods detail is essential. As seen in soil survey documentation in databases, detailed description of the analytical methodology used to measure survey parameters and the time frame for field sampling are needed (Lacarce et al., 2009). Staudigel et al. (2003) assert that any description of physical geological samples within journal publications should at minimum provide a full account of the sampling process and analytical process including "information on the origin of the data" and "how the data were normalized" (p. 7) though it is not a formal community practice. This emphasis on specific activities related to research data production further supports the role of methods for inclusion in metadata for long tail science data.

[3] Genbank, http://www.ncbi.nlm.nih.gov/genbank/

Scientists' metadata generation

The importance of detailing research methods as part of dataset metadata must be viewed in relation to researchers' practices of generating metadata for research data. The creation of metadata to enable sharing and reuse of data is generally met with resistance by scientists. Practice studies report a number of barriers to creating metadata commonly identified by scientists, including limited time, insufficient resources, and lack of incentives (Edwards, Mayernik, Batcheller, & Borgman, 2011). The level of contextual detail needed for understanding how the data was collected, the events of the actual collection, and any additional descriptions to ensure proper use by interested parties not only calls upon the expertise of the original scientist but may also involve information professionals who can help facilitate the creation of metadata and liaise with repository services for the storage and maintenance of the data over time (Baker & Bowker, 2007). However, for long tail science researchers who lack the resources for a designated information professional or data manager, documenting the research activities is often left to a graduate student or research staff who may not have the same degree of professional expertise (Mayernik, Batcheller, & Borgman, 2011). The responsibilities for creating the metadata may be delegated to a more expert laboratory field technician or a novice; or in other cases, this responsibility is never clearly assigned within a research team resulting in fragmented documentation (Mayernik, 2010).

Researchers are often not aware of discipline-based metadata standards or choose not to apply the standard due to the complexity or lack of technical support for implementation (Mayernik et al., 2011). Even minimum metadata requirements may not be satisfied. In the field of embedded network sensing, for example, Wallis, Mayernik, Borgman, and Pepe (2010) found limited compliance by scientists in submitting requisite metadata. Rather than the application of metadata standards, localized practices for description and documentation of data are more commonplace. Tenopir et al.'s (2011) survey of 1329 scientists from a variety of science disciplines—including biology and the environmental sciences and ecology— revealed that the majority (78%) either did not use metadata standards from their respective communities or applied local, home-grown practices for metadata. While there may be wide variation in what "local" practices constitute for research groups within a domain of study, sometimes these perceived differences are actually less local and more widespread. As observed by White (2010), scientists working in evolutionary biology discerned their practice of organizing research data as personal and unique yet similar styles of organization were found across

scientists, indicating that differences in practice may not be as pronounced. These local differences may be less of a barrier to facilitating comprehension of data if similar metadata practices are used.

In contrast to the local metadata practices of the long tail science communities, some "big science" enterprises, including astronomy and genomic biology, have relatively long established traditions of standardized data management practices, domain repositories, and data sharing, and remain exemplars in the design of curation infrastructure and services (Lynch, 2008). The increasing use of computationally intensive techniques for research by these disciplines is also reflected in technological advancements in tools and systems to record scientific workflows, which provides a semi-automated alternative to manually documenting the concise step-by-step description of the scientific procedure and protocols enacted (McPhillips, Bowers, Zinn, & Ludäscher, 2009). However, the use of these documenting technologies is not widespread across scientific domains (Davis et al., 2012) with little evidence from data practice studies to suggest workflow technologies for documenting methods as an emergent trend. There are some emerging efforts of integrating more standardized metadata generation practices into field-based sciences. For example, DataUp, a tool for managing and documenting tabular data incorporates the Ecological Metadata Language (EML) in metadata creation.[4] At present, the capture of methods metadata must continue to rely on techniques such as semi-structured interviews with data producers rather than automated output.

Another area where metadata generation for datasets plays an essential role for long tail science research is meta-analysis and synthesis studies. This type of research necessitates that data from different sources be well described in order to determine relevance for application to a new synthetic study, but also the various data sources must be interoperable. Meta-analysis is a statistical methodology to examine new research findings from independent studies, and the approach has grown in use, particularly in the Earth Science area of ecology, where there is "growing pressure on researchers to provide accurate quantitative assessments, predictions, and practical solutions to pressing environmental issues (e.g. biodiversity losses, biotic responses to global climate change)" (Koricheva, Gurevitch, & Mengersen, 2013, p. xi). Synthesis research is also a key driver for ecoinformatics, which "enables scientists to generate

[4] DataUp "Create Metadata," http://dataup.cdlib.org/dataup_features.html

new knowledge through innovative tools and approaches for discovering, managing, integrating, analyzing, visualizing and preserving relevant biological, environmental, and socioeconomic data and information" (Michener & Jones, 2012, p. 86). Similar interests in meta-analysis research are reported in other areas of Earth Science, including the soil science community, which looks to bring together data from in situ physical, chemical, biological (omics), and imaging techniques to better understand soil processes (Frontiers in Soil Science Research, 2009). Clearly, for meta-analysis research to flourish in the Earth Sciences, scientists will need to adherence to available scientific metadata standards and systematic provision of methods metadata.

Methods description and journal publications

An alternative approach to metadata generation for data by scientists is to supply a citation to the associated scholarly research publication. Data producers claim this "publication of record," which often describes facets of the dataset, the collection process or algorithms used to produce it, can be used in lieu of generating a metadata record for the dataset for public release (Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011). Similar observations from Earth Science data centers indicate journal articles are often included as citable references for a specific dataset, since these articles present a clear summary and analysis of the processes performed on the data (Parsons, Duerr, & Minster, 2010). From these findings, the details surrounding data production and the methods used stand at the forefront of what journal publications can provide in terms of metadata for the reuse of scientific data.

Scientists also use journal publications as sources of methods description when evaluating whether data are of sufficient quality for reuse. Within the long tail science of ecology, researchers utilize journal articles to identify timeframes or geographic elements as a means of systematic sampling when selecting data to reuse from multiple colleagues (Zimmerman 2007). The use of journal articles by scientists as a descriptive source for data also spans disciplines. Faniel and Jacobsen's (2010) study of data reuse by earthquake engineers details the critical role that journal publications have in conveying necessary methodological information for assessing the relevancy of the underlying data for use. The level of expertise and proficiency reflected in the description of experimental procedures along with accompanying tables and diagrams of the experimental set up and specimen properties contributed to reuse decisions. Journal articles are perceived by scientists as polished,

summarized analyses of the data yet remain a complement to the primary documentation recorded by the data producer (i.e. laboratory notebook) through the research process. However, ready access to scientists' documentation may be difficult to procure, as notes taken in the course of research are often recorded for personal use and may not be readily shared with others (Campbell et al., 2002), making journal articles a reasonable source to consult for methods description.

The importance of methods explication for data is seen in the emergence of data journals. Within the Earth Sciences, data journals have fostered publication of research datasets by adapting established journal publishing infrastructures. The concentration of data articles contained in these journal is on the description of a dataset and details related to its collection, processing, software, and file formats rather than presenting formal analysis and findings as with conventional journals (Chavan & Penev, 2011). The objective of the data article is to understand *how, when,* and *why* the data were collected and *what form* of the data is available for use. For example, the Geoscience Data Journal[5] from the PREPARDE project[6] is a new partnership started in 2013 between Wiley-Blackwell publishing and established Earth Science data repositories in the UK and across Europe. The purpose of this partnership is to develop appropriate workflows that allow data papers to be published in the journal highlighting a dataset. Similarly, the Earth Systems Sciences Data Journal,[7] launched in 2009, is an example of this new platform for open access data that emphasizes the role of methods in how data are collected and analyzed. It encourages authors to contribute reviews and examples of different methodological approaches that generate high quality data. While it is unclear if data producers conducting Earth Science research actively use these data publication venues, the emphasis on methods discussion within the data article content further illuminates the significance of research methods to understand data. The focus on research journal publications for this study to generate methods metadata can have potential application to populating data article content.

[5] Geoscience Data Journal: http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060
[6] PREPARDE project: http://www2.le.ac.uk/projects/preparde
[7] Earth Systems Sciences Data Journal: http://www.earth-system-science-data.net/home.html

Research design implications for encoding "methods metadata"

The discussion of methods used in the scientific research process is a primary contribution to generating metadata for data from the Earth Sciences. Returning to the research questions outlined, the first question addresses what methods metadata can be derived from semi-structured interviews as compared to content from journal publications. The first part of this question attends to research practices of data production (gathering, processing, analysis) as a starting pointing for identifying methods metadata. The second part posits how methods metadata can be obtained from different information sources.

The examination of research practice studies reveals the important role that research methods descriptions have for scientists to understand the processes and instrumentation used to gather and process data to facilitate data reuse. Research data practices are socially mediated and these interactions are an important part of data sharing, use, and management in research (Birnholtz & Bietz, 2003). Interviewing techniques to reveal data practices are common across these studies and are effective in gaining an in-depth perspective from researchers that can shed light on the social elements important for documenting methods in metadata. The interviewing technique also follows the tradition of qualitative studies of science, which have examined large-scale collaborative scientific enterprises (i.e. Galison, 1997) but have the potential to be applied to the localized research of long tail sciences. Journal publications are recognized in the literature as a source used by scientists not only as a way to understand the accompanying data but also as a surrogate for a dataset's metadata record. Given their active use by scientists, journal publications present a primary source for encoding methods metadata. By establishing the motivation through the review of literature related to the first research question for the application of semi-structured interviews and secondary analysis of journal publications to understand methods metadata, I next discuss the literature connected with differences in subdiscipline data practices and the role of metadata standards for scientific data.

Determining subdiscipline differences for methods metadata

The second area of inquiry builds on findings from data practice studies discussed above to look at how methods metadata differs across subdisciplines in Earth Science. In situating methods metadata within the curation context, the term "designated community" from the OAIS reference model provides a way to conceptualize the subdiscipline as the

targeted audience for whom the data are shared; this community should readily understand the metadata on methods made available with the data. As Parsons and Duerr (2005) indicate, the initial designated community might be part of a closely related scientific discipline as the research group producing the data, but since there is the potential for use by other communities of users, accompanying documentation must facilitate broad but appropriate use. Curation for data needs to be constantly evolving to accommodate the dynamic nature of knowledge base configurations and shifts in use from the original designated community. In this respect, the methods metadata needs are likely to differ from one subdiscipline to the next due to the basic contrasts in community knowledge and practices. Studies of scientists' practices, as discussed in the previous section, have provided input on the key role that research methods description has in assessing the quality of data for future use. Another important consideration is to what extent information on methods used to generate data, when provided by researchers, is represented in metadata standards for Earth Science and how these differences can be understood within scientific research communities.

Domain metadata standards and application for 'methods'

The use of standards in research is integral to transforming local knowledge into public knowledge (Zimmerman, 2003). Metadata is one area that has benefitted from community development for standardization (Yarmey & Baker, 2013) to support discovery of data; accommodate reuse by the original investigator and external researchers; and enable human and automated use of data (Michener, 2006). Standards for metadata are often rooted in the scientific practice of the community (Willis, Greenberg, & White, 2012). In the long tail sciences where data are heterogeneous and there is a tendency to use localized conventions developed within the small lab group environment (Wallis, Rolando, & Borgman, 2013), implementing a metadata standard can ease some of the variation that deters integration and reuse of research data. In this section, I review metadata schemes from several Earth Science-related fields to understand how description of research methods is captured.

"Methods" as a metadata element is represented in different ways within metadata schemes for data in the Earth Sciences. Table 1 illustrates examples of established metadata scheme for scientific datasets and areas within the schemes that provide a representation of

methods.[8] The sample of metadata schemes was adapted from Willis, Greenberg, and White's (2012) study on metadata goals for scientific metadata standards and was extended with data-oriented schemes that have potential application for long tail data ingest into a repository. The majority of these metadata schemes are applicable to Earth Science data with DDI and mmCIF as additional schemes for social science data and biomolecular data, respectively, that are well established in their respective fields and provide insight into methods representation that may not be covered by Earth Science-related metadata schemes. As shown in Table 1, the metadata schemes were examined to understand what and how information about methods would need to be included for generating metadata. Each metadata scheme was assessed based on the following criteria: Presence (confirmation that there is explicit use of "methods" or a similar derivation as an element name within the scheme); Detail (encompasses the information that should be included for a "methods" element or alternative elements that would contain methods-related description); and Status (denotes whether the "methods" element is required in the metadata scheme).

[8] An initial version of this table and the following analysis are presented in Chao (2014).

Table 1: Examples of metadata schema for datasets and representation of "methods" in schema.

Scheme name	Presence	Detail	Status
Content Standard for Digital Geospatial Metadata (CSDGM) *https://www.fgdc.gov/metadata/csdgm*	NO	**Lineage** - information about the events, parameters, and source data, which constructed the data set, and information about the responsible parties.	Mandatory with optional elements
Darwin Core (DwC) *http://rs.tdwg.org/dwc/terms/simple/index.htm*	NO	**samplingProtocol** - The name of, reference to, or description of the method or protocol used during an Event. Refer to <Event> for additional terms. <Location> A spatial region or place, named or not.	Recommended
Data Documentation Initiative (DDI) *http://www.ddialliance.org/specification*	YES	**Methodology** - concerns data collection, determining the timing and repetition patterns for data collection, and sampling procedures; content elements include: **DataCollectionSoftware**, **SamplingProcedure**, **DataCollectionMethodology**, **TimeMethod**, **DeviationFromSampleDesign**. Related: **SamplingProcedureType** - describes the type of sample, sample design and drawing the sample; supports the use of a brief term or controlled vocabulary to classify the type of sampling procedure described.	Mandatory
Directory Interchange Format (DIF) *http://gcmd.gsfc.nasa.gov/add/difguide/index.html*	NO	**Summary** - a brief description of the data set along with the purpose of the data, may include "scientific methodology"; **instrument (sensor name)** - name of the instrument used to acquire the data; <temporal coverage> specifies the start and stop dates during which the data were collected.	<summary> is mandatory; others are highly recommended
Dublin Core – Dryad Application Profile (DCDryad) *http://wiki.datadryad.org/Metadata_Profile*	NO	**dcterms:description** - Human-readable description of the resource; an abstract or summary. **Spatial Coverage** - description of the data specified by a geographic description and/or geographic coordinates. **Temporal Coverage** - description of the data, as geologic timespan.	Mandatory
Ecological Metadata Language (EML) *https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/./index.html*	YES	**eml-methods module** - describes the methods followed in the creation of the dataset, including description of field, laboratory and processing steps, sampling methods and units, quality control procedures; **eml-protocol module** specific to prescribed procedures	Not clear; suggests use of EML modules as needed
ISO 19115 *https://geo-ide.noaa.gov/wiki/index.php?title=ISO_Lineage*	NO	**Lineage** - based on *sources* which are either used or produced in a series of *process steps;* includes elements such as **LE_Source**, **LE_Algorithm**, **LE_Processing**, and **LE_ProcessStepReport**	Mandatory
Macromolecular Crystallographic Information File (mmCIF) *http://www.rcsb.org/pdb/static.do?p=file_formats/mmcif/index.html*	YES	**method_list** - part of Data category with related Data items; method description: stores the experimental method used to create the structure	Not mandatory

Among the Earth Science-related schemes, EML was one of the more robust schemes for articulating definitions and criteria specific to the processes of producing data. It was the only scheme examined with explicit metadata for methods description spanning processing steps, sampling methods, and quality control procedures. The Lineage element from CSDGM and ISO19115 contains similarities in definitions and structure with the methods metadata from EML and may also be a term that has greater significance within the geospatial data community than "method." "Lineage" is part of the Data Quality Information section of CSDGM and ISO19115, corroborating findings from Faniel and Jacobsen's (2010) study of data reuse by earthquake engineers on the importance of data production process documentation as a measure of quality for data.

Another metadata scheme without a visible element for methods description is the Dublin Core application profile developed by the Dryad repository to support the production of metadata for publications and the associated heterogeneous ecological and evolutionary biology data. While description of methods for collecting data is not explicitly covered by the properties in the Data Package Module of the profile, there is potential for this description within the scheme. For instance, Wallis et al. (2010) expand the "description" property from Dublin Core to accommodate information such as research questions, what variables were collected, data collection process and equipment, along with size and format (p. 337) for application in embedded network sensor research in ecology. Interestingly, both the EML (Michener & Jones, 2012; Leinfelder et al., 2010) and the Dryad Dublin Core application profile (Greenberg, White, Carrier, & Scherle, 2009) have been investigated for their use in automated metadata generation and the field of "description" (Dryad) and the "methodological information" module (EML) are still manually generated. One important aim of this investigation of developing methods metadata is to inform more automated approaches for identifying and extracting scientific methods to address this gap in metadata generation. These emerging efforts toward more automated metadata generation demonstrate techniques that can be adapted for gathering methods information.

Across these Earth Science-related schemes, it was not necessarily evident that methods metadata elements were required for the metadata record. Schemes such as CSDGM, DCDryad, and DIF have some required elements that may include methods information but methods description is generally not required. The optional use of these metadata elements for methods description reflects varied support for capturing metadata about methods. This also suggests

different aims or priorities in metadata generation for datasets, with other elements prioritized to satisfy a record for discovery and access.

Beyond the Earth Sciences, the description of methods is clearly explained in DDI, which is rooted in the social science data community. The inclusion of sub-elements in the scheme to further elaborate on the methods parallels the structure and level of detail displayed in EML. In contrast, the <methods list> element for mmCIF had minimal definition on what information to include about the data production process. These metadata schemes generally have some element(s) where description of methods could be included but identifying those element(s) that could be used for methods description was less straightforward than explicit use of "methods" as an element name. With this challenge of understanding and locating metadata elements from formal schemes, scientists' limited utilization of metadata standards (Tenopir et al., 2011; Mayernik et al., 2011) may be warranted if basic information about the method of data production cannot be readily documented or shared. The number of associated metadata elements specific to methods, as seen in the more robust schemes like EML, may equally overwhelm a researcher and also result in minimal compliance.

While the DDI, EML, and CSDGM, have more robust schemes for articulating definitions and criteria specific to the processes of producing data, one element that seems to span multiple disciplines and schemes is "sampling," a research technique generally applied for capturing a quantity or portion for analysis that is representative of a greater phenomenon. The appearance of "samplingprotocol" from the DwC often applied to biodiversity-related data and "sampProc" or sampling procedure from DDI used for social science-oriented data reveal different interpretations of what information should be encoded for these elements. For instance, <sampProc> (DDI) is centered on information about the data collection process with emphasis on sample size and sampling design description while <samplingprotocol> (DwC) is much more broad in definition encompassing any procedures used for sampling. The inclusion of sampling as a metadata element provides a more granular level of detail for describing the process of how data are produced.

The initial review of metadata schemes for scientific data provides evidence in support of methods description with the level of support varying across different disciplines. The recognition of methods as a descriptive component in these metadata schemes further confirms the significance of research on generating methods metadata. A more extensive examination of metadata schemes used for Earth Science data is part of the analysis of this

dissertation to determine how methods metadata can be adapted to existing schemes. With anticipated differences in metadata schemes, the resulting methods metadata is also expected to reflect the diversity of research from the Earth Science community,

Studying disciplinary differences

Studies of data practices across multiple disciplines offer another perspective on potential differences in methods metadata. Identifying how different scientific research cultures influence data practices is critical since "a generic approach to data curation will not be sufficient to cope with the different data-related needs and expectations of researchers working in different disciplines other than at a superficial level" (Key Perspectives Ltd., 2010, p. 2). Multi-domain or multi-discipline studies feature similarities and differences across a range of practices and attitudes toward data sharing, sometimes drawing from research fields across the sciences, social sciences, and humanities. However, analysis within these expansive categories and other broad disciplinary groupings can conceal critical distinctions within research or subject areas. The review of data practice studies featuring multiple disciplines and research fields makes visible the level of analysis (i.e. discipline, sub-discipline) to examine methods metadata differences.

The PARSE.Insight (2009) survey on researcher perspectives towards preservation and sharing of digital research data provides some insight into the complexity of disciplines. Respondents could identify with one of nine discipline categories and were allowed to specify a subject area with the category. Across the nine discipline categories, there were up to eleven subject areas identified for a single discipline, demonstrating the breadth of unique communities within a single discipline. There is also potential for more in-depth investigation to better understand the range of perspectives toward digital research data activity. Tenopir et al.'s (2011) survey of scientists' data sharing practices and expectations specified "biology" from the life sciences and "environmental sciences & ecology" from the natural sciences. However, other results from their study are reported at a more general domain level, such as "physical sciences" or "social sciences," making it difficult to make direct comparisons across these different research areas. While discipline-level differences were evident, these studies also reveal the potential for differences in data practices and expectations within a single research area.

The complexity and diversity of long tail science research makes analysis of an individual research area necessary to the expose multiple facets of practice essential for describing methods. In a comparison of research data sharing and communication across eight subject disciplines, Swan and Brown (2008) studied disciplines that spanned the spectrum of data sharing traditions and established infrastructures for data. Ten to fifteen experts were interviewed from each discipline providing a range of responses on emergent issues for a robust comparison across cases. For instance, the data descriptions within the most specialized research domain studied—the Rural Economy and Land Use (RELU) case—ranged from very informal, ad hoc practices to more formal standard metadata practice with projects that have a dedicated data manager. The attention to a single research area in this study not only demonstrated how different description practices were but also successfully incorporated the research context and external motivations that might influence practice, such as the inclusion of a data management plan required of all government-funded RELU research.

The focus and formation of cases also influence how differences in data practice can be more appropriately described for methods metadata. The DCC SCARP project conducted a collection of seven case studies that comprise researcher attitudes and approaches to data deposit, sharing and reuse, curation and preservation from research fields within engineering, architecture, atmospheric sciences, neuroimaging, medicine, and social sciences. Despite the immersive approach to develop each case, the authors relate,

> "…the case studies illustrate that the discipline is too broad a level to understand data curation practices or requirements. The diversity of data types, working methods, curation practices and content skills found even within specialised domains means that requirements should be defined at this or even a finer-grained level, such as the research group" (Lyon, Rusbridge, Neilson, & Whyte, 2010, p.4).

It is clear that the level of analysis for these discipline-based studies is critical to accurately inform data curation practice in accordance with researcher needs. As Cragin, Palmer, Carlson, and Witt (2010) suggest in their study on data practices in long tail science, the "subdiscipline" is the ideal lens to view nuanced practices from the scientists' perspective while also integrating local communication and the contextual elements that make practices possible. By focusing on this level of analysis, more concrete and adaptable findings on practices can be generated to inform description of data for curation.

The review of literature on multi-discipline data practice studies indicates differences not only at the discipline level but even more subtle differences visible within a discipline that would need to be accounted for in methods metadata. The examination of formal metadata schemes for scientific research data also situates methods metadata in existing structures for description of data by confirming the inclusion of methods description within these schemes. Building on the anticipated differences in methods metadata at the subdiscipline level within the Earth Science community, I further examine the literature on data sharing practices to inform what information about methods should be available when providing access to research data.

Data sharing insights for methods metadata

Studies of sharing data can contribute to further understanding of methods information to include for dataset metadata. Scientists' attitudes toward sharing their original research data with others beyond the local research team and collaborators are generally positive but usually with conditions in place regarding access (Tenopir et al., 2011). Within empirical studies of data sharing practices situated in the long tail sciences, scientists are reported to share based on the type of data requested, the individual requesting access, and when in the research process a request for data is made (Borgman, Wallis, & Enyedy, 2007; Borgman, 2012). Moreover, the format and accompanying documentation or how data are shared provides a tangible marker of the practice. Approaches to sharing data are an area in flux given contemporary funding requirements for open access to data and journal publisher attention to data availability. This also has an influence on the inclusion of methods metadata, which must likewise comply with new standards and guidelines for research data.

Dimensions of data sharing

Scientists are more likely to share certain types of data based on the approach used to gather the data and the status of the data within the research process. In a study of habitat ecology researchers, Borgman, Wallis, & Enyedy (2007) report scientists were more inclined to share data generated by automatic systems such as embedded sensors rather than those manually collected, a process that is time and labor intensive, and usually done with limited resources. Data that have just been collected or gathered, or "raw" data, are not usually made openly available, with researchers preferring, for practical but also cultural reasons, to release

a cleaned or processed version of the data (Swan & Brown, 2008). The raw data may be too large or in a proprietary format that prohibits them to be shared effectively. Some scientific communities have established conventions for what data are shared. Crystallography researchers, for instance, tend to only share processed data using the community-developed data format. These distinctions in what data are shared by scientists in different research fields are taken into consideration for this dissertation research; depending on what data are likely to be shared, appropriate methods metadata will need to be supplied to show how that data were generated.

Even with increasing expectations for open sharing of research data, researchers have concerns about who will have access to their data and how they will be used. There are apprehensions about data misuse, especially for scientists working in field locations that are ecologically sensitive or scientifically significant. Cragin et al. (2010) report concerns from a soil scientist about unsolicited industry use and that scientists who have experienced misinterpretation of their shared data believe it may reflect poorly on them as the original data collector. The Long-term Ecological Research (LTER) data repository takes precautions against misuse by allowing the principal investigator to stipulate that an interested researcher must contact and request the data as a condition of use; this process usually requires that the requesting researcher provide a detailed description of why and how the original dataset will be used (Karasti & Baker, 2008). Having this mechanism in place not only reveals who is interested in the data but also provides a channel for fostering collaborations, since common or complementary interests may be identified through interaction. There is also the potential for research data to be purposefully withheld. In the biological life sciences, data withholding behaviors are more common when produced as part of industry collaborations and also occurs as a means to protect a scientific lead prior to publication. The latter reason was the response for nearly 50% of respondents that reported refusing to share research materials and data with other university scientists (Blumenthal et al., 1997; 2006).

Determinations of when data will be made available are contingent on the research lifecycle. The general consensus for researchers in the long tail sciences is to specify restricted access to data until publication of results and findings, which can be several years after the research project has concluded (Cragin et al., 2010). Especially for the small research team, the time frame provides an opportunity for graduates students, post-docs, or research staff to publish before data are publicly released. Sometimes, the time frame for sharing data is already

established by a governing body. Scientists engaged in research at one of the ecologically diverse sites within LTER Network, for instance, are expected to deposit non-restricted data within two years from when the data were originally collected.[9] As part of this policy, the metadata must be made available regardless of restrictions on dataset access and include information on methods, structure, and quality assurance. With the funding agency requirements for data management plans and open access to data produced through grants, enforcement could include withholding funding if a plan is not implemented. It is not clear if or how often restrictions on funding has occurred or whether this is an effective mechanism that directly influences timely data sharing. In these instances where data sharing and access are delayed or restricted, the availability of methods metadata for the dataset could provide some information on what data were produced and the techniques applied. Having access to the methods could fuel new research even if the data may not be available.

When data sharing does occur, researchers in the long tail sciences rely on informal, personal communication. Scientists sharing data beyond the research group or collaborators often provide access based on personal request, which may involve several hours of preparation to locate and describe the dataset before sending (Cragin et al., 2010; Swan & Brown, 2008). Personal networks are also important in the discovery of new data resources and used by researchers in climate science and oceanography to request data (Chao, Cragin, & Palmer, 2015). These examples align with findings from social science researchers where informal sharing of primary research data is more common (44.6%) than formal sharing through an institutional repository or data archive (11.5%) (Pienta, Alter, & Lyle, 2010). Other scientific communities, such as astronomy, have formalized conventions for providing public access to data via well-established data archives (Key Perspectives Ltd, 2010).

Formal data sharing is often tied to publishing norms in research areas. One aspect of the emerging shifts in journal publisher requirements is the condition that articles for publication submission must make available the underlying datasets from the research. Some publishers will accept these datasets as supplemental materials and provide access to them through the journal website though this practices is still nascent (Swan & Brown, 2008). In other instances, a specific repository is designated by the journal for data deposit. Such a partnership is evident with a group of journals on ecology and evolutionary biology, which

[9] LTER data policy: http://www.lternet.edu/policies/data-access

have developed explicit policies on sharing and archiving data, and the Data Dryad repository[10], which hosts and links deposited data with the journal article (Whitlock, McPeek, Rausher, Rieseberg, & Moore, 2010).

Data sharing in journal publications

The "sharing" of data within the content of journal publications is connected with access to the data used in the research. As noted above, datasets can be accessed as supplemental materials to the published article made available by the journal, and this has been shown to significantly increase citation rates to the papers in a study in the area of microarray clinical trials (Piwowar, Day, & Fridsma, 2007). In a parallel study of journal articles from phylogenetic, earth and environmental, and genetic sciences, shared data were also frequently represented through data citations in the reference list, description of the dataset in the full text of the article, or as part of a footnote or in a table (Piwowar, Carlson, & Vision, 2011). The use of journal articles to convey information on where to access data can be thought of as metadata for describing an associated dataset.

Other approaches to sharing in research articles consist of unique identifiers or accession numbers assigned to a dataset by a repository or weblinks to departmental or institutional websites to access the dataset, as seen in an analysis of crop science journal publications by Williams (2012). This analysis also confirmed subfield differences in data sharing for the crop sciences. As would be expected, GenBank[11] was commonly used to provide access to data for articles on genetics-oriented studies. For articles based solely on field research,[12] only two (n=124) provided access to the data, although it was not indicated by the author what sharing approach was used. While these studies reinforce the prevalence of subfield differences, it is evident that sharing data is still nascent in areas, such as fieldwork studies. It is also possible co-authorship may suggest that data were shared among distributed authors but this cannot be determined unless somehow made explicit within the content of the article (Swan & Brown, 2008).

[10] Data Dryad repository site: http://datadryad.org/

[11] GenBank repository: http://www.ncbi.nlm.nih.gov/genbank/

[12] The article also distinguishes "field & genetic," "field & greenhouse/laboratory," and "field & greenhouse/laboratory & genetic" as separate categories of research publications.

In relation to methods metadata, sharing practices among researchers in the long tail sciences not only contribute to what type of data may be made available but also the temporal and social context of these transfers. When and with whom data are shared may predicate when methods metadata should be provided and refined. Data curation support to update and maintain the methods metadata will be necessary as reuse of data occurs over time after public release and the needs of potential user communities will need to be accommodated. The provision of methods metadata can potentially deter misuse of data, as discussed earlier, by making clear how the data were generated and subsequently used for research analysis. Instances of data sharing by scientists remain difficult to account for due to the informal and private nature of requests, but journal publications are a potentially fruitful source for discerning this practice and will become more so as public access to data increases along with data citation. As evidenced by studies of data sharing practices, the methods applied in producing data—the focus of this research—are a key part of both sharing and reusing data.

Conceptual foundations

The primary conceptual foundations for this investigation come from social studies of science. They have guided analysis and interpretation of practices that inform methods metadata in different subdisciplines of Earth Science. Latour's (1999) concept, *circulating reference,* provides an analytical frame for understanding the data production process and what methods metadata can be derived from interviews and journal publications. The second concept, *metadata frictions,* developed by Edwards and colleagues (2010; 2011), provides grounding for interpreting curation needs and responsibilities for methods metadata.

The concept of *circulating reference* was first introduced in the context of field research in the Amazon where botanists and pedologists recorded observations of the natural world to address the project goal of reporting on the relationship between savanna and forest environs (Latour, 1999). Since then, this concept has also been applied in science education research, particularly in training students to work in the field (Kastans, Agrawal, & Liben, 2009). *Circulating reference* refers to the active processes, or series of systematic transformations that lead to knowledge formation. These transformations can either amplify or reduce the content of the data by emphasizing particularity and continuity (amplify), or removing information in order to be more compatible and standardized (reduce). Evidence of these transformations is known as *references* or *referents.* With the accumulation of multiple references, this *chain of*

references is akin to the provenance of the data— starting from point of origin (e.g., gathering of the data from a field site) and moving through the series of transformations to the current form.

Within this research, the circulating reference concept can be applied to analysis of the data production processes and what information about the data is being amplified or reduced. For instance, a researcher may reduce observations about a study site to present in a related table that conforms to publisher standards for a journal article. The table is now a referent and as an individual sees it within the journal publication and requests more information about how the table was generated from the data producer, amplification occurs because new inquiry is being fostered. Viewing the data production and reuse process as a series of transformations helps in discerning the stages where methods metadata will be of importance in documenting amplification or reduction. Without methods metadata as a reference, the potential knowledge embedded in the data gathered cannot be realized.

The second concept of *metadata frictions* is applied to facilitate articulation of how this research can be applied for curation purposes. Frictions arise during the activities involved in creating, handling, and managing metadata products, which Edwards et al. (2011) frame as "metadata-as-process." Within the context of their study of atmospheric science modeling, metadata production as a process is often manual and of ad hoc quality, and can be further characterized as fragmented (i.e. contributions by many individuals); divergent (i.e. different versions of metadata can be created); and iterative (i.e. reconcile versions, overcome miscommunication). The metadata product is also oriented toward privileging local use over more open sharing with the community. The frictions stem from tensions and costs in time, money, energy and attention needed to produce metadata. The different types of frictions are further elaborated by Mayernik et al. (2011) in their examination of three cases of large, distributed, collaborative science projects where frictions are identified for standardization, time/temporal, data sharing, and human support availability in creating and managing metadata. Across both studies, frictions in the production of metadata that arise in scientific collaborations never seem to be resolved but can, in part, be mitigated through informal communication between the scientist who produced the data and fellow collaborators.

Metadata frictions can occur at different stages of the research process and involve not only scientists but also data managers or other information professionals. By looking at the process of identifying and collecting methods metadata for data, potential areas where frictions

may arise become more prominent. The metadata frictions that arise from standardization, for example, may not be as pronounced within long tail science research where scientists adhere to local practices to document data, but for data curators, the absence of standards gives rise to frictions for ensuring future discovery and use of data not only by the original researcher but others. In this respect, my study seeks to identify strategies to reduce those frictions in metadata production by examining different approaches, both direct and unobtrusive, to attain methods metadata for scientific research data.

Background literature: Summary

To enable the reuse of data, practice studies have demonstrated the importance of research methods to explain the context of how data are produced by scientists and for inclusion as methods metadata for data. I draw from these studies to position my first set of research questions.

A) What *methods metadata* relating to how data are gathered, processed, and analyzed for research can be derived from:
    - qualitative interviews?
    - research journal articles?

These practices, related to the production of data in the research process, are only part of what constitutes metadata for a dataset but are nevertheless an important component to prioritize. This research question is not aimed at producing an exhaustive account of methods metadata but rather at understanding what sources can be used in the development of metadata for research data with the emphasis on the study methods. Other products generated during the research process by scientists such as personal notebooks could also contribute to identifying methods metadata but remain an area for future investigation, especially as documentation intended for personal use may be fragmented and not technically sound for immediate sharing (Orlikowski, 1995). The novel use of journal articles as a source for methods metadata does not necessarily supplant the need to speak directly with scientists about the procedures undertaken to generate and analyze their data. Both the analysis of journal article content and interviewing of data producers can contribute to new techniques for metadata generation.

Differences in metadata are anticipated at the subdiscipline level of research where scientists in long tail research employ customized methods in working with data. While *methods* are currently covered by metadata schema for the Earth Sciences, there is limited guidance on what information about methods (i.e. cited protocols, spatial or temporal setting for data gathering, etc.) should be included. In formulating the second research question, the level of analysis at the subdiscipline level is a change from most previous practice studies, which have tended toward comparative analysis at the discipline level. Examining methods metadata differences exhibited at the subdiscipline level within the Earth Sciences can bring a more refined perspective on the nuances of subdiscipline practices in relation to data. The second question is:

B) How does methods metadata differ across subdisciplines in Earth Science?

By examining how scientists discuss the activities surrounding data production in research through semi-structured qualitative interviews and within formal journal publications, this dissertation addresses gaps in how metadata for research data can be generated by focusing on the research methods of long tail scientists as critical information for data sharing and reuse.

CHAPTER 3: RESEARCH DESIGN

This dissertation uses a case study approach to investigate the identification and recording of research methods for metadata purposes. As discussed in the preceding chapter, the case study method has been effectively used for practice studies. In particular, analysis at the subdiscipline level has been successful for distinguishing data practice differences, allowing "critical focus on the domain questions and data types that produce the 'science' in a community—the social unit where data sharing practices and reuse can best be explored" (Cragin, Chao, & Palmer, 2011, p.441). Three cases were developed for subdisciplines within the Earth Sciences: Soil Ecology, Volcanology, and Stratigraphy. These areas of Earth Science were useful for comparison of practices since scientists across all three apply similar empirical approaches for collecting observational data from field sites for analysis and experimentation.

The strength of case study design is the opportunity to gain an in-depth perspective into a particular phenomenon based on evidence from multiple sources and techniques (Yin, 2009). In developing the subdiscipline cases, I collected different types of evidence from participating scientists on their data production practices and activities by applying two different qualitative methods—semi-structured interviewing technique and content analysis of journal publications. The analysis of practices and activities is aimed at determining methods metadata for research data from each subdiscipline.

Case studies have also been widely employed in studies of information behavior including the identification of differences in the information practices of scholars within the digital realm (Fry, 2006); activities, strategies, and behaviors employed by interdisciplinary information seekers (Foster, 2004); and the influence of research culture on e-resource use (Talja, Vakkari, Fry, & Wouters, 2007). Together, these studies demonstrate the flexibility and variation in what constitutes a "case" and how the case study approach is effective for framing comparative analysis. The case study development for this research is not only used to examine the complex nature of research data production and analysis within and across subdiscipline communities but also to provide insights into the efficacy of the two different sources, interviews and journal articles, for identifying methods descriptions for datasets.

As seen in the literature reviewed, other approaches have been used to investigate scientists' research practices with data. These techniques include online surveys, which have the advantage of providing responses on attitudes and behaviors from a larger population of scientists with the potential for cross-discipline comparison (e.g. Tenopir et al., 2011).

However as noted in the preceding chapter, the level of detail on practices at the subdiscipline level is difficult to attain. Even with the inclusion of open-ended survey questions for participants to elaborate on practice activities, responses "often do not produce useful data" and may be incomplete, lack clarity, or be inconsistent with the question objectives (Fowler, 2009, p. 72). Another approach would be the use of ethnographic methods (e.g. Borgman, Wallis, & Enyedy, 2007), which allow for prolonged engagement with the science research groups to more fully understand the cultural and social norms and practices. Although techniques such as qualitative interviews are used in ethnographic work, the embedded nature of ethnographic research was not feasible for this study and would be too great of a time investment compared with the qualitative interviews for the work of data curators.

Among the three cases, Soil Ecology was the most extensively developed. This research area was of particular interest since data reuse, especially in the form of meta-analysis, is a growing trend in Soil Ecology. Meta-analyses are used to understand patterns in findings across existing individual experiments that can potentially generalize to the field at large. Improving and supporting data for meta-analysis research in Soil Ecology requires reconciling the wide variation in how different studies report data compilation (Hungate et al., 2009). The variation in reporting stems from differing levels of detail relayed about the data source and related gathering and analysis processes, which emphasizes the overarching need for more consistent metadata if data are to be shared and reused.

Enhancing description about the production and analysis processes for data would provide greater access to information necessary to conduct meta-analysis research. While data description to support meta-analysis was analyzed as part of the Soil Ecology case, all three cases examined scientists' application of techniques and standards for the research activities related to producing and analyzing data. Including the Volcanology and Stratigraphy cases allowed for comparison across subfields, and also contributed to the transferability of findings to the broader Earth Sciences.

Overview of research design

The study was conducted in three phases (see Figure 1 for research design overview).

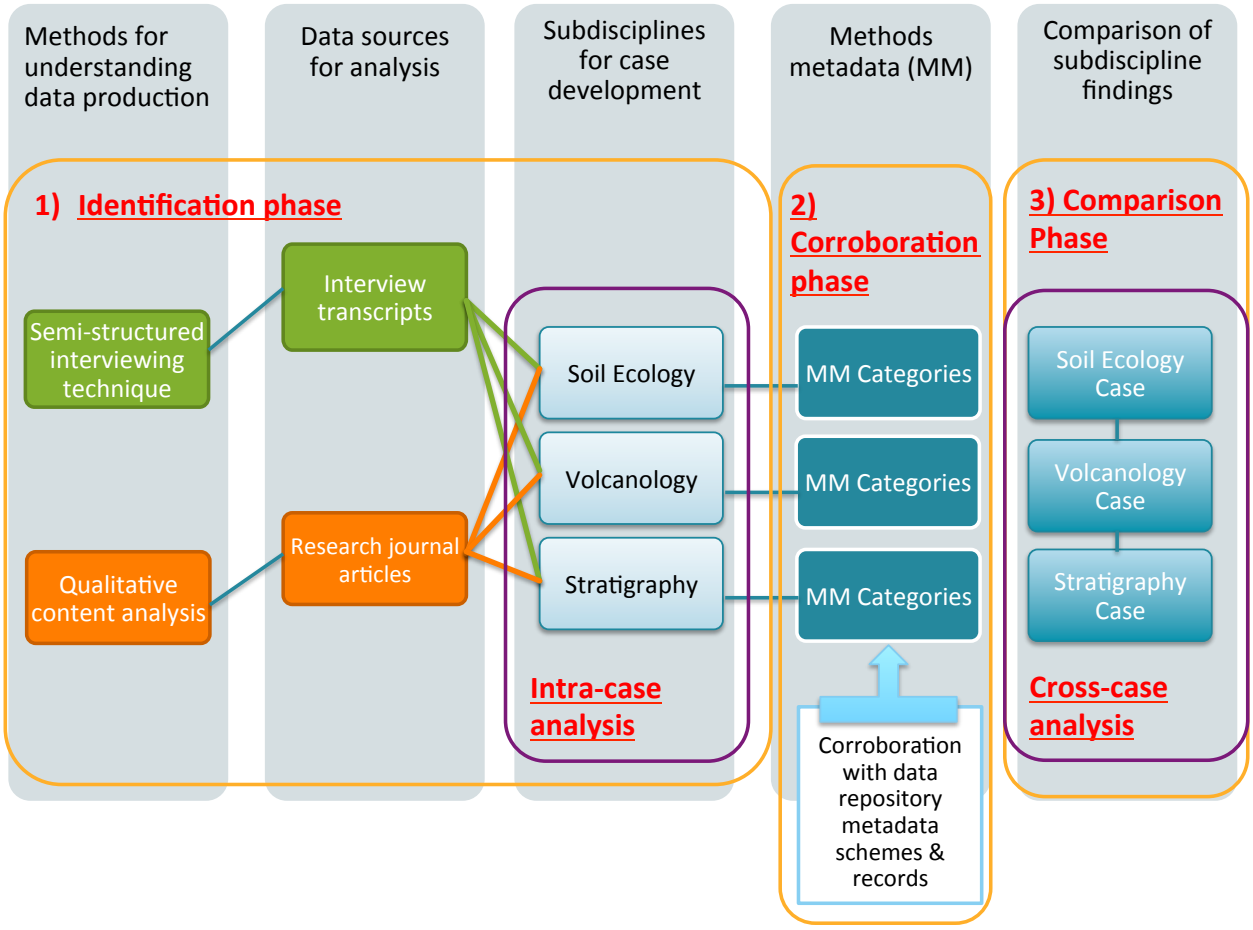

Figure 1: Overview of the research design detailing the three phases of research and related analysis.

Phase 1 focused on identifying methods metadata for data based on two sources of evidence: 1) qualitative interview transcripts and 2) research articles from journal publications. The qualitative interview data for the three identified subdisciplines were gathered from the Data Conservancy Data Practices research (Cragin, Chao, & Palmer, 2011). While some analysis of the interview data has been completed by the Data Practices team with the larger data set that includes the Soil Ecology, Volcanology, and Stratigraphy sources (see Weber et al., 2012), none of the previous work examined these subdisciplines in depth or specifically investigated the role of data practices in determining metadata description for data. The qualitative data collected from the Data Conservancy project was coded and analyzed anew for the specific aims of this study.

Although the interviews conducted were not specifically aimed at understanding metadata generation by scientists for a particular dataset, gathered data were relevant for this study because they are illustrative of the kind of narrative description of data processes that interviews generate. It was presumed that multiple interviews and regular interaction with the data producer would result in a comprehensive account of how a dataset was generated and what information should be included in the metadata record. The objective of looking at these interviews in relation to journal publications was to investigate the pros and cons of a non-intrusive approach to securing description about research methods from scientists.

The bulk of the peer-reviewed journal publications sample analyzed were authored by the interview participants and supplemented with additional articles from respective discipline journals to validate findings from participant publications. A pilot study was conducted to understand the viability of using publications to determine methods metadata for data. The findings of the pilot study are discussed in a later section of this chapter.

The second component of Phase 1 was the intra-case analysis for each subdiscipline. This systematic analysis brought together the evidence from direct engagement with scientists through interviewing and the unobtrusive approach of examining scientists' research publications to understand what information regarding methods for producing data can be established for each subdiscipline. Subdiscipline findings are synthesized into Methods Metadata Categories (MM Categories).

Phase 2 of this study corroborated the synthesized findings from the Phase 1 analysis for each subdiscipline case. The MM Categories were compared with respective subdiscipline data repository metadata records and associated metadata schemes to determine how robust the derived methods metadata is and what gaps may still exist in methods description.

Phase 3 built on the findings from the first two phases to address what differences exist for identified methods metadata information across the subdiscipline cases. This cross-case analysis incorporated the data repository metadata record assessment with the MM Categories development for comparison of the findings from the subdiscipline cases within the broader context of the Earth Sciences. Each of the three phases is discussed in greater detail in the following sections.

Phase 1: Methods metadata identification

In the following account of Phase 1, I detail the processes employed to generate the data products from a semi-structured interviewing technique and discuss how journal publications were gathered for content analysis. I also describe how these two data sources were used in the intra-case analysis for each subdiscipline in the identification of methods metadata information for research data.

A) Qualitative semi-structured interviews

Background

The foundation for each case consisted of qualitative data gathered as part of the Data Conservancy, a NSF DataNet cyberinfrastructure project led by Johns Hopkins University. The Data Conservancy drew expertise from scientific research communities, digital library engineers, and the information professions to design and construct tools and services for the long-term preservation, sharing, and cross-disciplinary discovery of research data. As a Data Conservancy partner, the University of Illinois had two research teams examining different aspects of curation requirements: Data Practices and Data Concepts. The Data Practices team conducted comparative qualitative research on data production and use practices of scientists within the Earth and Life Sciences in relation to curation needs, cultures of sharing, and re-use potential of data (see Appendix C for publications). The Data Concepts team addressed identity conditions and representation levels for datasets through conceptual model development (e.g. Renear, Sacchi, & Wickett, 2010). The Data Conservancy project began in Fall 2009 and concluded in July 2012.

As a member of the Data Practices team, I had an active role in the recruitment of participants, design of the data collection protocol, and interviewing process. A total of 28 interviews with 20 scientists were completed. Fifteen interviews from 10 scientists are used for this study; I conducted 10 of the interviews used for this study either independently or in partnership with Dr. Melissa Cragin. Dr. Cragin conducted the remaining 5 interviews. I have access to all the interview transcripts, worksheets, and other artifacts collected during the interview process. For the interviews I participated in, I also have observational notes taken during the sessions.

IRB approval

The involvement of human subject participants required approval of the study by the University of Illinois' Institutional Review Board (IRB). The package of materials submitted to the IRB by the Data Practices team consisted of the study design, recruitment email and letter, participant consent forms, Pre-Interview Worksheet and interview protocol (see Appendix A for IRB documentation and approval letter, and Appendix B for approved instruments). Participation in the study was voluntary and participants were free to withdraw from the study at any time. The research protocols employed posed minimal risk to participants. Since the completion of the Data Conservancy project, no participants have withdrawn from the study.

Interview participant sampling

As detailed in Chao, Cragin, and Palmer (2015), the participant recruitment process drew on three purposeful sampling strategies – *criterion*, *theory-based*, and *maximum variation* (Patton, 1990). The initial *criterion* was to identify and recruit participants representative of the service communities needing data services like those provided by the Data Conservancy. The *theory-based* strategy was aimed at identifying research areas that are examples of "long tail" disciplines, which have research data considered to be in high need of dedicated curation services due to variable data management practices, inconsistent application of standards, and heterogeneity and complexity in the data types produced (Heidorn, 2008). Finally, *maximum variation* in participants was sought for the sample. The diversity of perspectives from researchers in similar disciplinary communities contributes a more robust understanding of research practices and their complexities compared with a single viewpoint within a community.

The initial recruitment of participants was from Data Conservancy partner organizations consisting mainly of academic research universities. The Earth and associated life sciences research areas were the original fields identified for service, with a particular focus on soil science, geology, and computational modeling for earth systems phenomena. Key informants from the initial recruitment of participants from Data Conservancy partner organizations—individuals who provided rich description on personal research but also a community perspective on research practices—were asked to recommend additional researchers who may be interested and appropriate for taking part in the study. This "snowball" technique for recruiting can be more effective since recommendations are more

likely to result in participation compared with a cold contact approach (Creswell, 2012). The snowball technique also helped to focus recruiting within the research community, which aided in developing a more comprehensive case study. Key informants and recommended participants were contacted via email and asked to participate in the study (see Appendix A for recruitment email and letter).

In order to achieve maximum variation in conjunction with the snowball technique, faculty, researchers, staff, postdoctoral scholars and graduate students were included in the sample to reflect the range of responsible parties actively involved in the research process. This purposeful sampling also aimed to include group members (i.e. faculty, researchers, staff, postdoctoral scholars and graduate students) within the single research team to assess reliability of observations on research data production and analysis processes and sharing of data from individuals exposed to the same local conditions. Multiple research teams from comparable research areas within the Earth and life sciences were targeted within the same institution and across Data Conservancy partner organizations in order to understand the potential differences in research data practices within a subdiscipline. For each of the three selected subdisciplines, there was at least one research team represented with interviews from multiple members. The Soil Ecology case, for instance, included three research teams from two different institutions.

Interviews were conducted with (5) Soil Ecology participants, (3) Volcanology participants, and (2) Stratigraphy participants, yielding a total of (10) participants (see Table 2 for overview of participants). In addition to the interviews, the other data produced from the qualitative interviewing technique included written responses to questions in the Pre-interview worksheet, observation notes, photographs and artifacts of the research process from participants. Review of the interview data for this study confirmed coverage of local standards for documenting the data production process, including the use of different reference guides and books consulted for research design development and classification of physical samples. It was also evident that scientists informally share information about a particular data collection technique or study site through personal requests. In general, the interviews with scientists offered insight into the purpose of, and rationale behind, their research conducted within the greater scientific context and the types of data collected. These preliminary observations support the importance of understanding research practices and

their contribution to identifying data production methods metadata for data through qualitative interviewing.

Table 2: Summary of participant sample for each subdiscipline.

Case: Subdiscipline (# of participants)	Participant ID	# of interviews collected	Role	Research specialty
Soil Ecology (5)	C1SE1PI1	2	Principal Investigator	Soil geobiology
	C1SE1GS1	1	Grad student	Soil management
	C1SE1GS2	1	Grad student	Earthworm systematics, molecular phylogeny and soil ecology
	P1SS1PD1	3	Post Doc	Soil microbiology
	P1SE1GS1	1	Grad student	Soil microbial ecology
Volcanology (3)	C1GEO1PI1	1	Principal Investigator	Petrology and volcanology
	C1GEO1PD1	2	Post Doc	Volcanology and magma dynamics
	C1GEO1GS1	1	Graduate student	Magma systems
Stratigraphy (2)	C1GEO3PI1	2	Principal Investigator	Sedimentology
	C1GEO3PD1	2	Post Doc	Geological time scales

Process of semi-structured interviewing technique

The semi-structured interviewing technique for Data Conservancy was a multi-step process generating a number of data products for analysis.[13] The data collection instruments were pre-tested with two scientists and refined prior to implementation on the targeted research population. The process began with a *Pre-Interview Worksheet* (Worksheet), a brief questionnaire orienting participants to the objectives of the study and the scope of questions that will be asked during the Research Interview. Worksheet responses provided insight to the

[13] See Cragin, Chao, & Palmer, 2011, for further details on the conceptualization of the Data Practices research design; see Chao, Cragin, & Palmer, 2015, for application of the qualitative approach; see Appendix B for examples of data collection instruments.

domain science and the types of data involved in the conduct of research. Participants were also asked on the Worksheet to recommend research publications representative of the research conducted in his or her area; these publications were reviewed in preparation for the interview. The Worksheet was followed by a semi-structured *Research Interview*, which focused on questions regarding the generation, use, and management of various types of data most important for access and use by others internal and external to the research area. The majority of interviews occurred in-person with remaining interviews taking place over the telephone.

With selected participants, *Follow-up Interviews* were conducted to address any gaps from the Research Interview but also to go more in depth on an emergent issue or topic. Based on the Research and Follow-up Interviews, these participants were recognized as having potential data contributions that could be accommodated for ingest to the Data Conservancy repository. *Data Interviews* were also carried out as part of the follow-up process to document specific details about data types with potential value for reuse by other scientists and what requirements would be necessary to facilitate the ingest of these data into a repository. Finally, *Lab Visits* were conducted and often resulted in the collection of research data artifacts, particularly sample data sets and associated documentation generated by the scientist along with photographs and field notes captured by the Data Practices team. In practice, the objectives and activities associated with the Data Interviews and Lab Visits were often achieved during Follow-up Interviews rather than as separate occurrences.

B) Journal articles

Overview

The second data source was gathered from research journal articles authored by the interview participants. Examining articles added an unobtrusive dimension to the cases that offered further information on specific studies conducted by participants and the opportunity to assess how research methods and data production processes may align or vary within a subdiscipline as reported in formal publications. Additional articles beyond those of the participants were added for each case as a point for internal validation in order to examine how the larger research community discusses data production and to what extent variations exist within a community. The journal articles used in this study differ from those publications

reviewed as part of the pre-interview process from the Data Conservancy research, which were read for general background and context of the scientific study.

Pilot study findings for journal article content analysis

A pilot study was conducted to assess the viability of using peer-reviewed research publications for identifying descriptive information related to data. I anticipated from my experience with the Data Conservancy Data Practices research that the processes for generating and processing data for analysis would be well documented in the journal publications due to the rigorous recording practices instilled in field-based research. For instance, interviewed scientists described that the level of detail in publications should be sufficient to reproduce a study. The sample (n=15) consisted of five publications from each subdiscipline selected from interview participant curriculum vitae (CV), which included peer-reviewed journal articles, conference proceedings, book chapters, and reports. I targeted different kinds of publications to ascertain their inclusion of data-related information. In this pilot study, content analysis was conducted using inductive reasoning (Corbin & Strauss, 2008).

The primary findings of the pilot study on scholarly publications are as follows:

- Finding 1: Reports, book chapters, and conference proceedings did not provide enough information related to data production and analysis processes to be used effectively for methods metadata.
- Finding 2: Within journal articles, the "Methods" section was the most information-rich regarding data production and analysis processes.

In the first finding, the content information provided in these different types of scholarly publications varied in regards to methods description. For instance, conference proceedings tended to be limited by word counts and the provision of methods used in a study was brief, if discussed at all. The absence of methods description was also evident in retrieved reports and book chapters where the subject focus did not warrant this type of description. In addition to the inconsistent reporting of methods in these publications, the actual retrieval of papers also posed challenges as many of the citations listed in CVs and personal webpages were not made publically available. The most reliable sources were research journal articles, which discussed methods and were also relatively more accessible than the other kinds of scholarly publications due to consistent indexing in databases.

The analysis of the journal articles resulted in the second finding concerning the information provided in the "Methods" section of a paper. Descriptions from this section included information about the data that were collected and how they were processed and analyzed. The ease of locating this information within articles differed across the subdisciplines. Figures and diagrams were prevalent, showing study site layouts and the mapping of areas where data were collected. These visuals complemented description of research practices within the narrative text of the article.

Relative to the "Methods" section, other sections of the articles relayed minimal information about the data or implemented methods. It was anticipated that the "Discussion" or "Conclusion" sections would describe the limitations of the research and address issues related to the data production process. However, none of the sample publications communicated such issues. The "acknowledgements" section included information on who participated in the data production. For instance, one Soil Ecology paper recognized the students who participated in soil sample preparation. Such information brings a level of transparency to the data production process. Overall, the pilot study findings demonstrated the significance of journal article sections in conveying information related to data production and distinguishing specific research practices in connection with how data are produced, verifying the usefulness of journal publications as a data source for identifying methods metadata related to research data.

Journal article sampling

The sample of journal articles for analysis in this dissertation was developed from several sources. The initial sample of peer-reviewed publications was drawn from the literature produced by the interview participants in the three subdiscipline areas. This purposive sample was used to complement the interviews and related products collected in the first part of this study in order to reveal how descriptions of the research methods relayed by participants in interviews matched description in journal articles. Utilizing a purposive sampling approach is characteristic of qualitative content analysis whereas the sample for quantitative analysis would need to be developed for statistical inference (Zhang & Wildemuth, 2009). Participant CVs, research webpages, Google Scholar,[14] and Research Gate[15] were used to

[14] Google Scholar "About," https://scholar.google.com/intl/en-US/scholar/about.html

distinguish the number of journal articles published over the research career of the scientist through 2013. For the article sample from interview participants, non-English language papers, reviews and editorials from the journal and research articles that did not report methods were excluded. Based on the pilot study described above, other kinds of publications (i.e. editorials, book chapters, conference proceedings) were excluded as they yielded less robust information regarding research practice and data description and were often difficult to access and retrieve. The sample was further refined to only include articles in which the participant was the "corresponding author." This designation typically connotes the person who is the first point of contact and is knowledgeable about the article to respond to inquiries. The designation of "first" or "last" author to mark lead authorship differs across scholarly communities and their publication practices and therefore was not used in this study.

An overview of the articles retrieved from interview participants is summarized in Table 3. Distinctions were made between the number of articles the interview participant was the lead author on and the actual number of articles that were included in the final sample. It is important to note that the final number is primarily composed of lead author articles. Additional non-lead author articles from participants were included in order to be more inclusive of those participants that did not have as much academic experience as others (and may not have had a lead author paper).

Common research themes and topics also emerged from participant journal articles and were used in expanding the sample of articles for analysis. The focus of articles from Soil Ecology was on studies of biological invasive species as represented by earthworm communities in urban and experimental field sites. Research areas from Volcanology research papers included approaches to studying crystal size distribution in different geological formations and the contributions of these approaches in the development of magmatic models. Lastly, themes from participant articles in Stratigraphy covered the astronomical time scale used to map the history of the Earth's movement in relation to changes in climate and sedimentary systems with particular interest in Milankovitch cycles. These themes were then used for expanding the sample for content analysis and identifying metadata records from data repositories in Phase 2 of this study.

---

[15] Research Gate "About us," http://www.researchgate.net/about

Table 3: Summary of journal articles collected from participants and included in the sample for analysis.

Case: Subdiscipline (# of participants)	Participant ID	# of journal articles identified	# of articles as lead author	# of articles included in sample	time span of collected articles
Soil Ecology (5)	C1SE1PI1	23	7	7	1985-2013
	C1SE1GS1	1	0	1	2013
	C1SE1GS2	5	3	5	2004-2012
	P1SS1PD1	2	2	2	2009-2013
	P1SE1GS1	5	3	5	2008-2012
Total # of Soil Ecology participant articles				**20**	
Volcanology (3)	C1GEO1PI1	18	1	3	1981-2012
	C1GEO1PD1	1	0	1	2010
	C1GEO1GS1	1	0	1	2011
Total # of Volcanology participant articles				**5**	
Stratigraphy (2)	C1GEO3PI1	34	3	7	1987-2012
	C1GEO3PD1	5	4	5	2000-2011
Total # of Stratigraphy participant articles				**12**	

This initial sample was expanded with additional articles related to each of the subdisciplines from highly cited journals. These journals were determined using the Scientific Journal Rankings (SJR[16]) with the following subject categories applied to identify journals: "Soil Science," "Stratigraphy," and "Geochemistry and Petrology." The selection of articles from these journals was determined based on common research themes identified from participant articles and used as search terms to discover related texts. Finally, only articles published between (2006-2013) were used in the final sample. While a longitudinal perspective would bring attention to the potential development of methods description over time, more recent articles were used in this study to complement the time period when the majority of articles by interview participants were published.

Based on these extracted search parameters, I used Web of Science[17] to identify and retrieve related journal articles. Selected articles had to include some description of methods and not just a theoretical discussion. The final sample was refined to encompass an even

[16] SJR, part of Scimago Journal & Country Rank, which is a "portal that includes the journals and country scientific indicators developed from the information contained in the Scopus® database (Elsevier B.V.)." (SCImago. (2007). SJR — SCImago Journal & Country Rank. Retrieved from http://www.scimagojr.com). Journals are ranked through the SJR Journal Rank Indicator, which measures the impact, influence, and prestige of a journal by taking the average number of weighted citations received in a selected year by the number of documents published in the journal in the past 3 years. The rankings used for this study are taken from the year 2012.

[17] Web of Science "About," http://wokinfo.com/citationconnection/

distribution of articles for each year in the specified time range (see Table 4 for overview of additional articles). For articles referencing an earlier publication for research methods, the earlier or original publication was integrated into the sample, regardless of publication year. As the objective of this research is to inform metadata generation for research data, it was advantageous to have access to the greatest amount of detail on these activities to be included in methods description. There were two instances where new articles were introduced to the to sample: (1) for Volcanology and (1) for Stratigraphy. The number of publications analyzed for each case study varied, with the aim of reaching saturation—a threshold where further data collection does not reveal new findings (Corbin & Strauss, 2008) about the manner in which data are presented and discussed.

Table 4: Summary of additional articles for the content analysis samples for each subdiscipline.

	Soil Ecology	Volcanology	Stratigraphy
Journals for sample	• Applied Soil Ecology • European Journal of Soil Biology • Soil Science Society of America Journal • Plant and Soil	• Estuarine Coastal and Shelf Science • Geology • Earth and Planetary Science Letters • Geochemistry, Geophysics, Geosystems (G^3)	• Contributions to Mineralogy and Petrology • Geosphere • Journal of Petrology • Journal of Geophysical Research
Research theme search terms	"earthworm"; "invasive species"; "meta-analysis" & "soil"	"crystal size distribution"; "meta-analysis"; "whole rock chemistry"	"multi-taper method" (MTM); "cyclostratigraphy"; "meta-analysis"; "Milankovitch"
# of articles retrieved	18	10	10
# of participant articles	20	5	12
Total # of articles	**38**	**15**	**22**

Process of qualitative content analysis

I utilized a qualitative content analysis approach to systematically identify themes from the documents related to descriptions of data production and analysis practices that would be important to capture for data curation. This qualitative approach allowed for the range of meanings of the methods practices to be revealed and was more conducive for this research than a quantitative approach, where content is viewed objectively to determine frequencies of

instances in the text (White & Marsh, 2006). As discussed by Zhang and Wildemuth (2009), qualitative content analysis is an established approach in information and library science research providing rich description of particular settings or phenomena; this description can then be used in developing each case study. Each paper was examined for information directly related to how and what data were gathered, processed, and analyzed; the processes and procedures executed including tools and instruments used; and indication of data sharing or reuse from external sources. The framework for analysis is discussed in detail in the following section.

Intra-case analysis

Analysis was conducted on the data originating from the qualitative interviews and journal publications for each subdiscipline. The primary focus of analysis was scientists' practices related to the generation and analysis of data—specifically, the statements they make about their practices applicable for methods metadata information for data. It is important to note that this analysis is not an evaluation of how content analysis and interviews are conducted, but rather how the processes for generating data are described by the scientist in-person and in-print and what each account can contribute to the cultivation of methods metadata. The findings for each subdiscipline also provide a basis for comparing methods metadata across the three research areas addressed in Phase 3 of this research.

In the first part of the analysis process, both the journal articles and qualitative interview data were manually coded using the same analytical framework. From the data generated by the semi-structured interviewing technique, the semi-structured interviews provided the most direct sources for analysis of practices. The additional data products— self-report written responses to questions in the Pre-interview worksheet, observation field notes and artifacts of the research process collected from participants—were used to better understand the context of the research practice in a subdiscipline.

The initial analytical framework applied in the coding of the interview data and published literature was developed based on the research questions, emergent themes from the pilot study on journal publication content analysis, and the Data Practices and Curation Vocabulary (DPCVocab). The DPCVocab was developed in part by using qualitative studies with Earth and life scientists to map relationships between data, research practices, and curation functions (Chao, Cragin, & Palmer, 2015). Although the vocabulary was not developed as a tool

for coding, all three of the vocabulary's categories—Research Data Practices, Data, and Curation—have terms that are effective for guiding systematic analysis of the interviews and journal articles. For instance, the main category on Research Data Practices provided a foundation of terms to elaborate on data producers' "gathering," "processing," and "analyzing" practices as conveyed in the research questions of this study. Additional terms from the vocabulary bring attention to data producer practices in relation to the data generated: terms such as *data characteristics*, *data stages, and standards use* from the Data category can situate the metadata generation process with terms from the Curation category, for example, *representation, stakeholders and responsible parties*, and *ingest*.

An initial coding of a small set of interview data and journal articles from each subdiscipline was conducted to test the usefulness of the code list. New concepts that emerged from the analysis of the different data sources from each subdiscipline were added to the code list, which was refined accordingly. I re-coded the initial set of interview data and journal articles to align with the other data. Atlas.ti (v 6.3),[18] a qualitative analysis software program, was used for coding and tracking versions of the code list.

In the second part of the analysis process, the following concepts and themes were assessed based on the coded qualitative interview data and journal articles from each subdiscipline:

- prevalent data types, data parameters/variables, and what procedures or protocols were used;
- standard practices used for data collection, processing, or analysis that conform to research community norms (such as use of a named reference guide or citation);
- evidence of localized practice for data description that does not align with research community norms (i.e. modifications to standard practices);
- attribution of responsible parties for the data production process;
- other emergent themes from coding

Throughout the analysis process, I also recorded memos to capture findings, observations, and interpretations for use in constructing each case.

[18] Atlas.ti homepage, http://atlasti.com/

Methods Metadata (MM) Categories development

A key outcome of the subdiscipline intra-case analysis is the formation of MM Categories, which were corroborated in Phase 2 of this study. The set of categories consist of terms identified from the intra-case analysis that detail practices connected with gathering, processing, and analyzing data. The MM Categories were tailored for each subdiscipline case in order to capture any distinct nuances in methods description from the analysis. The development of the MM Categories to represent methods metadata themes aligns with the qualitative analysis technique of designing *data displays* to organize and assemble information in a compact form for immediate accessibility (Miles & Huberman, 1994).

The set of MM Categories was then applied to journal articles from the sample set to test the efficacy of terms for identifying methods description from research papers in the different subdisciplines. Each journal article was individually assessed for content related to the set of categories. The number of journal articles containing content that could be extracted for each category was determined for the subdisciplines. The journal article frequencies were determined for each MM Category and provided support for analysis of subdiscipline differences in methods metadata.

Phase 2: Corroboration of methods metadata

In Phase 2, the MM Categories were verified through the examination of existing metadata records for scientific data and their associated metadata schemes. The initial findings on methods metadata from Phase 1 were compared to metadata records from relevant domain data repositories to evaluate completeness and verify the applicability of categories. The structured form of the metadata record provided a baseline for what methods description fit or could be added to existing records, particularly those metadata fields related to the data production and analysis process. In addition, content from metadata records illustrated current practice for how methods are documented and represented for existing datasets.

I examined records from data repositories in the Earth Sciences complementary to each subdiscipline. The DCC Disciplinary Metadata site[19] was used to determine established metadata standards applied in the Earth Sciences and the data repositories in which they were applied. Additional repositories were identified from the participant interviews and journal

[19] DCC Disciplinary Metadata – Earth Science, http://www.dcc.ac.uk/resources/subject-areas/earth-science

article content. The data repositories used for analysis in Phase 2 are listed and described in Table 5.

Table 5: Overview of selected data repositories for metadata records analysis.

Knowledge Network for Biodiversity (KNB) *Website:* http://knb.ecoinformatics.org/ *Metadata standard*: Ecological Metadata Language (EML) *Overview*: Internationally sourced repository intended to facilitate ecological and environmental research. Supported by the National Science Foundation. Over 21,000 data packages.
PANGAEA *Website:* http://www.pangaea.de *Metadata standard:* repository-based standard, compatible with ISO19115, Directory Interchange Format (DIF), and Dublin Core. *Overview*: Used by Earth Science Systems journal and other publications as a data repository. Online in 1995, PANGAEA is hosted by the Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven and the Center for Marine Environmental Sciences (MARUM), Bremen in Germany. (# repository holdings unknown).
EarthChem Portal (EarthChem) *Website*: http://www.earthchem.org/portal *Metadata standard:* repository-based standard *Overview*: A "one-stop-shop" for geochemistry data of the solid earth from multiple data systems. The portal returns integrated search results from the federated databases PetDB, SedDB, GEOROC, NavDat, USGS, and GANSEKI. Established in 2003 and hosted at Lamont-Doherty Earth Observatory. Over 800,000 samples deposited.
Global Change Master Directory (GCMD) *Website:* http://gcmd.nasa.gov/ *Metadata standard*: Directory Interchange Format (DIF) (transforms to CSDGM, ISO19115) *Overview*: One of the largest public metadata inventories in the world; maintains a complete catalog of all NASA's Earth Science data sets and services. Started in 1994. Over 31,000 datasets.
Geological Society of America – data repository (GSA) *Website*: http://www.geosociety.org/pubs/drpint.htm#aboutdrp *Metadata standard:* repository standard *Overview*: It is an open file in which authors of articles in GSA journals can place information that supplements and expands on their article. Established in 1974. (# repository holdings unknown).

Existing metadata records from repositories were selected based on the following factors:

- comparable scientific research themes revealed from the journal article sample for each subdiscipline.
- metadata records corresponding with data published between 2006-2013 to reflect contemporary metadata description practices for data.

The research theme search terms for each subdiscipline were queried in each repository. As detailed in Table 6, the search terms did not always yield records from a repository. This was

expected given the collection scope of the repositories. PANGAEA and GCMD contain data with broader coverage of Earth Science research whereas EarthChem, KNB and GSA are targeted to a subarea of Earth Science such as solid earth research or ecology. The GSA data repository was included in the sample due to the popularity of GSA journals among Stratigraphy study participants as a venue for publication. It is also a journal requirement for data producers to deposit the underlying data for a published research article with the GSA data repository. The overall variation in data repository holdings and description standards offers a more robust basis for verifying the set of MM Categories. Records retrieval occurred in October and November 2014.

Table 6: Summary of repository records retrieved for corroboration of MM Categories.

	Soil Ecology	Volcanology	Stratigraphy
Research theme search terms	"earthworm"; "invasive species"; "meta-analysis" and "soil"	"crystal size distribution"; "meta-analysis"; "whole rock chemistry"	"multi-taper method" (MTM); "cyclostratigraphy"; "meta-analysis"; "Milankovitch"
Repository and # of records retrieved			
PANGAEA	12	5	10
Global Change Master Directory (GCMD)	8	16	15
Knowledge Network for Biocomplexity (KNB)	12	0	0
EarthChem	0	12*	0
Geological Society of America (GSA)	0	0	8
Total # of records	**32**	**33**	**33**

The content of retrieved metadata records were analyzed in two areas:

1) Completion of methods metadata element(s)
2) Compliance of methods content with metadata element definition

* The terms "Journal of Petrology" and "Contributions to Mineralogy and Petrology" were used in the repository search since original search terms did not yield any results. These new search terms were derived from the journals used in retrieving articles for content analysis, only these two journals for the Volcanology sample produced viable records. Entries retrieved from the PetDB portal (single rock samples) came from 2 studies (J of Petrology: 35 entries; C to Mineralogy: 24 entries). Due to the repetition in methods content for the samples, only unique records were used for analysis.

For each data repository, any standards for metadata description applied by the repository were identified. The identified metadata schemes were either developed specifically for repository use or a known standard for Earth Science data. Methods-related elements were then isolated from the scheme; these elements were determined from review of available documentation about the metadata standard, specifically targeting elements with definitions related to data generation and analysis procedures (i.e. processes used to collect data from the field, etc.). See Table 9 in Chapter 5 for more details on identified elements from existing schemes.

Each metadata record was first assessed for the completion of methods-related elements. It was anticipated that not all possible methods-related elements would actually be used during metadata generation, especially if they were not considered mandatory or required for completing a record. Based on the availability of information contained in these methods-related elements, the second part of analysis examined the quality of methods information from the record. The element definitions or set of criteria for metadata description detailed from available documentation provided the basis for comparing content provided in these fields with the definition. For example, the definition for "Description" from the Ecological Metadata Language scheme is "The coverage field allows for a textual description of the specific sampling area, the sampling frequency (temporal boundaries, frequency of occurrence), and groups of living organisms sampled (taxonomic coverage)."[20] A metadata record entry for "Description" would be marked "low compliance" if information was provided for some but not all of the components listed (i.e. sampling area, sampling frequency, groups of living organisms sampled) in the definition. The examination of information quality demonstrated how expectations were met for metadata completion and what information, if any, was actually provided about the methods procedures.

The metadata records and formal schemes from each repository were then compared with the MM Categories. Similarities and differences in methods description from the repository records and schemes with the MM Categories were documented for each subdiscipline. I also utilized introductory domain science methods reference books to check

[20] Ecological Metadata Language "Description," https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/./eml-methods.html#description

methods terminology, and the National Environmental Methods Index (NEMI),[21] a database for analytical and field methods specific to environmental monitoring to corroborate the MM Categories. NEMI is focused on documenting research methods and it is a productive source to understand what information about methods to include and how methods can be organized to share with others. The consultation of methods reference books also provided a level of expertise in standard procedure descriptions of methods in the research setting to contrast with actual methods description supplied by scientists in scholarly research papers and in-person. The methods reference books were located based on a keyword search of the University of Illinois library databases for each subdiscipline. These reference materials were not necessarily subdiscipline specific but tended to be broader in scope. For instance, "geological field methods" were used for Volcanology and Stratigraphy and "soil methods" were utilized for Soil Ecology.

The products from this phase include annotated versions of the MM Categories and mappings with metadata schemes. Memos were also recorded to note the corroboration process and the role of interviews and journal articles as information sources to complete metadata records for each repository. The corroboration of the MM Categories for each subdiscipline contributes to Phase 3, which reviews subdiscipline differences in methods metadata.

[21] NEMI "About," https://www.nemi.gov/about/

Phase 3: Comparison of subdiscipline findings

The observations and findings for each subdiscipline from the first two phases of research are compiled for Phase 3 for cross case analysis. Each subdiscipline case is composed of the analysis of MM Categories application to journal articles and the analysis of MM Categories corroboration with available metadata records from data repositories and the associated formal metadata schemes. Based on the observations and findings compiled for each case, the cross case analysis focused on the following areas:

- the use of journal articles and MM Categories to identify and generate methods metadata.
- the role of semi-structured interviewing in methods metadata generation.
- the relationship between methods metadata and existing metadata schemes for data description, including metadata record population for data repositories.

The primary product of this phase was the synthesis of similarities and differences for identifying and populating methods metadata from interviews and journal articles across the three subdiscipline areas. The documented similarities in methods metadata were especially critical to capture since these observations could be more readily extended to the domain of Earth Science as a whole. Part of this synthesis included discussion of the adaptability of techniques used for identifying data production and analysis information to other field-based sciences to aid curation efforts in producing metadata.

Validation of subdiscipline findings

For each subdiscipline case, the validity of inferences and interpretations regarding methods metadata were established through the different phases of the research design. Phase 1 incorporated data collected by two different techniques (i.e. interviewing and content analysis) and encompassed perspectives from several scientists to build a more comprehensive understanding of the research practices within the subdiscipline and the potential range in variation of information related to description of methods. The use of multiple techniques to look at methods is meant to ameliorate the shortcomings of a single approach while also exemplifying their respective advantages (Guba, 1981). The internal validation of methods metadata at the subdiscipline level was established by utilizing the findings from qualitative interviews and the inclusion of practices and views from multiple scientists to confirm observations from content analysis of journal articles. Further external validation of methods

metadata was supported in Phase 2 through the analysis of identified methods metadata with existing metadata records and formal schemes for data, which represent real-world metadata structures that the identified metadata from this study could potentially be applied.

The third Phase of research encompassed the triangulation of methods metadata across the three subdisciplines in the Earth Sciences. Triangulation provides a mechanism to confirm that observations and findings carry the same meaning under different circumstances (Stake, 1995). Within Phase 3, I assessed how the findings from the Soil Ecology case, can generalize to the Earth Sciences by examining the methods metadata findings from Volcanology and Stratigraphy. Through investigation of potential differences in methods metadata at the subdiscipline level, I can better understand the prospective variation of methods metadata that would need to be accommodated in metadata creation for curating Earth Science data.

Data management

A variety of data were generated from this study. In Phase 1, the qualitative interviewing technique produced interview transcripts, audio files, worksheet response text, and documentary materials such as field notes and photographs with artifacts collected from some participants. In order to ensure the anonymity of participant identities, each individual was assigned a coded identifier that was applied to all documents for analysis pertaining to that person. Interviews from Phase 1 were digitally recorded and the audio files maintained for a minimum of three years from the date of creation on a secure server. The majority of interviews were transcribed by a third party and reviewed with the original audio file for typographical errors. While the transcriptions remained as faithful as possible to the original recording, frequent “umm”s or “OK”s were often omitted. Any field notes or artifacts collected from participants were also labeled with the coded identifier.

From the content analysis, resulting data consisted of digital copies of research publications in different formats (primarily PDF, but also HTML, and plain text). The research publications were labeled with a modified identifier retaining the same subdiscipline headings but the publication year was used in place of the data collection date. Journal articles with the same publication year are distinguished further with an alphanumeric code.

Data available from the corroboration phase encompassed the annotated metadata records from data repositories and associated scheme documentation. Each of the records was assigned an identifier corresponding with the subdiscipline. Information related to methods

description in the metadata schemes was compiled and organized in Excel spreadsheets. From both Phase 1 and 2, there was a set of annotated MM Categories for each subdiscipline stored as Excel spreadsheets. The analysis for Phase 3 drew on these spreadsheets for cross case investigation.

Limitations of case study design

One of the primary limitations of utilizing the case study approach was establishing the reliability of the coding performed on journal publications and transcripts. Without the benefit of multiple coders, interpreting interview transcripts and texts and applying appropriate categories is at personal discretion and not confirmed by inter-coder reliability measures, which aim to achieve high consistency among multiple coders. The inability to enroll additional coders can be attributed to the background knowledge needed, which includes a level of understanding of the data source content, coding framework, and metadata for the Earth Sciences. However, with the understanding that "the goal of reliability is to minimize the errors and biases in a study" (Yin, 2009, p. 45), I was diligent in documenting my research procedures in a manner that someone external to this study would be able to understand and potentially repeat the steps I have taken in designing these cases. In the scenario of coding, I utilized the features of the Atlas.ti software to comment on decisions made in assigning codes. These comments supplied important contextual information as I documented the steps of the study. This documentation takes the form of a detailed *case study protocol* (Yin, 2009) and contributes to the reliability of the coding.

Another caveat of qualitative research is potential participant bias during interviews and the accuracy of responses provided. Participants may be swayed to provide answers that reflect themselves in a more positive light or what they believe interviewers would like to hear (Creswell, 2012). In order to ameliorate these issues, I built a rapport with participants through multiple points of interaction established through the Data Conservancy research. I drew on my own experience in biology and social sciences research along with reviewing the literature recommended by participants to be indicative of their research field in order to build background knowledge on the subdisciplines. The references and links to publications in the field helped me to gain a broader understanding of the scientific discourse and better grasp the current research problems and approaches employed in the subdisciplines to guide my engagement with participants. My understanding of the science conducted in each

subdiscipline was further expanded with the focus on journal articles as part of first phase of research.

A third critique of case studies is their lack of generalizable findings. Case studies tend to portray a specific context that cannot fully represent similar contexts or situations. While the subdisciplines of Soil Ecology, Volcanology, and Stratigraphy are part of the Earth Sciences, it was not possible to statistically prove that the case study findings are representative of the Earth Sciences. Instead, I examined the compatibility of findings from these subdiscipline cases. As Foster (2004) describes, a characteristic quality of qualitative research findings is that they are "transferable" rather than "generalizable." In order to determine whether one set of findings will be applicable in another situation, an investigator must communicate to a reader how the original findings were derived and in what setting. There is greater emphasis on contextual and methodological description in qualitative studies and this description is included in the discussion of the research design and in Chapters 4 and 5.

CHAPTER 4: METHODS METADATA - PRACTICE & CONTEXT CATEGORIES

The primary objective for the first phase of research was to understand how to identify and organize methods description from content analysis of journal articles and interviews with science researchers. A key product from this first phase was the synthesis of findings from the different data sources to develop the Methods Metadata Categories (MM Categories). The MM Categories highlight the common factors used to describe research methods and data production and analyses processes specific to a subdiscipline. This set of categories can be applied and adapted as part of metadata infrastructure in data curation systems.

Development of Methods Metadata Categories

Formulated from the analysis of journal article content and interview transcripts, and the review of the DPCVocab, the set of MM Categories consist of two types of categories: "practices" and "context." The "practices" categories comprise information on the activities and set of procedures that typically express what actually took place in generating and analyzing the data for a research study. These practices are supported by "context" categories, which relay details surrounding the implementation of research practices such as the location or instruments used. The categories (noted below in italics) illuminate those research practices integral to understanding methods and also lend insight to contextual details used in these procedures. The definitions for all categories are detailed in Table 7. Together, this set of terms provides a more holistic picture of how methods represented in journal articles can potentially be included as metadata. Appendix D contains a summary table of the MM Categories derived from each of the data sources.

Table 7: Core set of Methods Metadata Categories, definitions, and category type. Relationships between categories are indicated in some of the definitions.

Categories	Definition	Type
Collecting	Process of gathering data, usually situated in a *study location*.	Practice
Reuse	Use of existing data (*data source*) for a new purpose/analysis.	Practice
Sampling	Procedure used to identify data (*data source*, *sample*) gathered for *analysis*.	Practice
Processing	Preparing data for analysis; can include any processes related to normalizing or homogenizing data for analysis, quality control measures, or handling of physical sample.	Practice
Analysis	Systematic process applied to data to answer proposed research questions.	Practice
Research scope	Research study context in which the method was applied.	Context
Study location	Physical location where collecting occurred, can include history or background of the location.	Context
Sample	Data product for analysis often collected from the *study location* (*collecting*).	Context
Variable/parameter	Characteristics gathered from a *sample* or *data source*.	Context
Data source	Existing data not collected by a data producer for the purposes of *Reuse.*	Context
Data access	Where and how data generated from research study is made available for use.	Context
Citation	Reference used in the description of a procedure (i.e. named technique), organization or institution, or another researcher that provided the data for analysis; can be applied with *instrument*, *software*, and *data source.*	Context
Modification	Explicit changes, adaptions, or deviations to a standard protocol/cited procedure.	Context
Instrument	A tool or device (manual or automated) used to implement a research procedure.	Context
Software	Program (proprietary or open source) used to assist in any part of the research process for data production.	Context

DPCVocab Contributions to MM Categories

The DPCVocab primarily contributed to the development of the "practices" categories. The "Research Data Practices" section of the DPCVocab, which relates to the research activities enacted in the conduct of the research study, was most useful in providing a base set of terms. The arrangement of the DPCVocab is structured to show relationships between different terms that add clarity to broader, overarching terms. For example, the term "processing" encompasses the terms "cleaning," "removing outliers," and "normalizing," while in the MM

Categories, only the term *processing* appears. The use of the more general "processing" affords retention of differences in terminology, especially as methods description varied across the Earth Science subdisciplines. Other contributions of the DPCVocab to the MM Categories include variations on "collecting," "sampling," "reusing existing data," and "analyzing."

Content analysis contributions

As a complement to the research practice terms contributed by the DPCVocab, the content from journal articles and interviews supplied contextual support for these practices in methods description. There were 10 categories identified for context to accompany the 5 practice categories. Below, I discuss each of the terms added to the MM Categories (see Table 7 for complete set and definitions).

The *research scope* category contains information to situate the overarching context in which data were produced. An example of category content includes available information found in journal articles about the objectives and questions addressed in the associated research study that can in turn inform why a particular method was selected. The *sample* generally refers to a physical entity directly collected by the scientists from a physical space or *study location.* Related to information about the location and sample are associated *variables* or *parameters*, which are the products of data analysis. Some studies utilize *data sources*, or existing data, as a supplement to *sample* data or as the sole data collected for analysis.

The description of particular procedures enacted during the data production and analysis process often included *citation* information in the form of a bibliographic reference. Citation provision is an established scholarly practice across scientific communities when publishing research journal articles (Brown, 2010). The citation of a method and related software and data sources can be evidence for verifying and asserting research expertise which in turn influences how quality of produced data is perceived by future researchers (Faniel & Jacobsen, 2010). Related to *citation* was recognition of *modifications* or explicit description of changes or adaptations to referenced procedures. Other context categories that emerged from interviews and journal articles include *instrument* and *software* use. These context categories helped to support description of standard or localized techniques based on the selected instrument or software program. For instance, one Soil Ecologist opted for remote sensing devices to collect real-time data on environmental conditions such as temperature and air moisture (C1SE1PI1_20100604). The use of remote devices differs from the more traditional

approaches of utilizing handheld devices or consulting the National Weather Service for gathering that environmental data. Providing information about the device used for data gathered also influences the likelihood of that data being shared and subsequently used by others (Borgman, Wallis, & Enyedy, 2007; Wallis, 2012).

A category that primarily emerged from interviews was *data access* or how data produced are made available. Participants across the three subdisciplines acknowledged an interest in making data produced available for public use, however in practice data mostly tend to be shared through personal networks and by request. From Volcanology and Stratigraphy interviews, physical samples such as whole rocks or thin section slides are shared between colleagues through personal requests. It is not clear, however, what methods documentation may be provided during these sharing instances. Interestingly, information for data deposited in a repository appeared in articles from two journals in the sample for Stratigraphy and one article from the Volcanology sample. Description of access to data may be an activity gradually taking root in the other Earth Science research areas.

The initial formation of the MM Categories centered on identifying methods description for each of the subdiscipline cases. While the actual research scope and questions differed from one study to another in a single subdiscipline, an underlying typology for information conveyed for methods emerged. The comparison of different information types for methods description across the three disciplines revealed that an all-encompassing set of terms could be generated to address the facets of data production and analysis. For this reason, a single overarching set of MM Categories was developed from the Earth Science subdiscipline cases. Unique features specific to the subdiscipline are included as "descriptive components" or subcategories to the core set of categories.

MM Categories: descriptive components

Some of the MM Categories are elaborated with subcategories to guide the kind of information needed for a particular category. The descriptive components were derived from the analysis of the interviews and articles and are tailored for each subdiscipline based on observed differences in data production and analysis procedures across the fields. These descriptive components were associated with each of the practice categories (*collecting, reuse, sampling, processing*, and *analysis*) and for *study location* and *instrument* from the context categories.

Of the MM Categories, descriptive components were identified from each subdiscipline for *sampling* and *study location*. For *sampling*, the main set of descriptors included "# of samples" and "sampling area" which were visible across subdiscipline fields. The "# of samples" was associated with the number of rock samples collected for Volcanology and Stratigraphy research. For Soil Ecology, a quantification of samples could usually be extrapolated from information related to the number of plots, blocks, and soil replicates taken for the sampling procedure. The "sampling area" can pertain to dimensions for blocks and plots in Soil Ecology or the range where samples are taken at designated intervals (i.e. samples taken at 150m intervals over 3km vertical spacing) as observed from Stratigraphy and Volcanology research. For Volcanology methods, the "sampling area" could also be represented as a geographic feature such as a quarry or road. In terms of subdiscipline differences for *sampling*, the "sample dimensions" can refer to precise measurements such as "25 mm diameter cores" (Stratigraphy) or relative size such as "the size of a softball" for rock samples (Volcanology). Specific to Soil Ecology was "sampling depth" or the distance downward (i.e. 20-30cm) that samples were extracted from a field location.

An identifying information feature to include for *study location* that spanned the subdisciplines was "geographic information." This descriptive component encompasses at least one of the following characteristics: longitude and latitude coordinates, full-text entries for country, city, or state, setting (i.e. urban, suburban, rural), or even the name of a research site. Other descriptions for *study location* were more generic such as an in-house or external laboratory to situate where a methods process transpired. Depending on the research, multiple *study location*(s) could be identified for different processes. For instance in Soil Ecology, the *study location* for *collecting* data typically occurs at a field site while the *study location* for *processing* that data normally takes place at the laboratory the data producer is based. Soil Ecology articles also described "soil type" information as part of *study location*, which would not typically be relevant for Volcanology or Stratigraphy metadata.

Differences in descriptive components for a single category were also determined. The *processing* of data in Volcanology acknowledged in journal article content highlighted the use of external services, such as laboratories with specialized equipment for analysis. The Volcanology interviewees confirmed external laboratory service use, particularly for processing whole rocks collected from the field, which results in data on chemical composition and thin section slides for further visual and microscopic investigation. Although the use of

these services is a recognized part of the methods process for Volcanology research, the actual procedures undertaken by these external laboratories are not explicitly documented or discussed in journal articles. It is also not clear whether researchers receive a copy of the procedures utilized by these services when data are returned from being processed. While obtaining the methods used by laboratory services would support more robust documentation of methods contributing to data production, documenting references to the service providers would support description on where data were processed.

Additional subdiscipline-specific descriptive components for MM Categories were observed for *analysis* and *collecting* in Soil Ecology. Descriptions of statistical tests conducted as part of the research study were consistently available from journal articles and would be identified as *analysis* metadata. The interviews with Soil Ecology researchers also emphasized aspects of statistical analysis, suggesting the value of documenting "variance" and the "confidence interval" used for the tests would provide greater understanding of the data. In the *collecting* category, the descriptive component of "date" relays a temporal information related to the season (e.g., spring) or the month(s) and year that collection took place. The unique features captured in descriptive components for methods description help to accommodate some of the distinctions in data production processes observed across the three subdisciplines.

MM Categories groupings

The relationship between practice and context categories influences their use in identifying methods metadata. With a practice-related category such as *collecting,* associated context categories would include what *sample* was taken, the *instrument* used for the sample extraction, as well as the *study location* where collection of samples occurred. For instance, *study location* description would distinguish if data were collected from a field site or in a laboratory environment. This distinction of *study location* in relation to data *collecting* was particularly helpful for descriptions of experimental studies. From the Soil Ecology analysis, both field sites and laboratory settings were used in conducting experiments while experimental methods were primarily concentrated to the controlled laboratory for Volcanology research.

Another practice-context category grouping is seen with *reuse* and the associated context category of *data source*. The *data source* may also have *citation* information to indicate the particular data repository from which the existing data was accessed. The groupings of

practice-context categories continue to grow when the *processing* and *analysis* of data are described; the addition of *sample*, *instrument*, *software*, *variable/parameter*, or other context categories to form a group can help more fully describe a data production or analysis process. Examples of practice-context category groupings for each subdiscipline are detailed in the following section (see Tables 8-10). Documenting the connections between the different types of categories is critical for robust description of methods for metadata inclusion.

Subdiscipline observations: Methods metadata from journal articles

I examined what information journal articles can contribute for methods metadata. I used the MM Categories as a framework to guide the review of each set of journal articles retrieved for the three subdisciplines. As previously described, the practice and context categories formed natural groupings for representing some of the data production and analysis procedures and these groupings were used as part of the review. I determined the number of journal articles containing description for *research scope*, *data access*, and the practice categories (*collecting*, *reuse, processing*, and *analysis*). Based on the number of articles identified for each of these categories, I further established how many of these articles contained information for context categories. For instance, 11 out of 20 journal articles may only have methods information on *collecting* data. The related context categories such as *sample* would then have values determined out of "11" articles rather than the total number of articles from the sample (i.e. "20"). Tables 8-10 show the values for the number of journal articles with methods description for metadata in each subdiscipline.

The initial analysis of Soil Ecology articles demonstrated the effectiveness of the MM Categories and groupings, thus the same categories and groupings were applied for Volcanology and Stratigraphy. Within Tables 8 through 10, the categories were coded (high, medium, low, none) to show the availability of methods information from the sampled journal articles. Those categories with methods description that were available from at least 70% of articles were coded as "high" (green), the middle range (30-70%) was designated as "medium" (yellow), less than 30% had the code of "low" (red), and no availability of methods information from the articles was coded "none" (white). Since the total number of articles differed for each subdiscipline, the ratios for were scaled accordingly. Findings from the Earth Science subdisciplines and discussion of journal articles as methods metadata sources are discussed in the following sections.

Soil Ecology

Overall, the set of journal articles from Soil Ecology was highly successful as a source for methods metadata (see Table 8). Description was consistently available from the 31 journal articles for the practices of *collecting*, *sampling*, *processing*, and *analysis* and for the majority of related context categories. The following is a typical example of the rich detail offered in journal article content:

> The sampling took place during a period with low precipitation in April 2007 (4.1 mm; measured at The Jena Experiment field site by the Max Planck Institute for Biogeochemistry, Jena). Normally, precipitation in April is about 27.5mm at the field site (mean of 2003–2006). Thus, the mean soil water content of the upper 15 cm was only 12% (mean field capacity of Ap-horizon 18% [1], Table 1). Four adjoined blocks were established parallel to the river to account for changes in soil abiotic conditions (Table 1) as a function of distance from the river [17]. At each block (ca. 60 by 280 m) we established 20 plots of 0.25m^2, spaced at 1m intervals, by removing carefully the upper 2–3 cm of the soil with a rake (80 plots in total). The removed topsoil was hand-sorted for earthworms and detected individuals (primarily epigeics, see below) from each plot were preserved alive in separate plastic bags filled with Jena soil (Eisenhauer, Straube, & Scheu, 2008, p. 142)

Information such as the "The Jena Experiment field site by the Max Planck Institute for Biogeochemistry, Jena" can be provided for *study location* and the detail of hand sorting for earthworms precludes the use of a manufactured device or *instrument* for this process.

Analysis description for Soil Ecology was predominately focused on statistics, which would lead to *software* rather than *instrument* information to accompany this type of analysis. Although it was not a frequent occurrence in the reviewed articles, the *reuse* of data was observed in studies utilizing a meta-analysis approach. Evidence of *modifications* to cited procedures or techniques was also infrequent, usually represented as a rationale given for a particular procedure used in order to overcome a limitation in the chosen technique.

While methods metadata could generally be identified from Soil Ecology journals articles, *data access* was the one MM Category for which information was not detected. This lack of information was consistent across multiple journals, suggesting that providing information to access generated research data was not a common practice in published papers. The Soil Ecology interviews confirmed that few researchers make data publically available. As one example, a Soil Ecology scientist was consulted on how a locally developed method could be adjusted and tailored for a specific soil type; it was later revealed that sharing and seeking

advice on methods was commonplace in the soils community more so than research data (P1SS1PD1_20100511).

Table 8: Availability of methods metadata from journal articles in Soil Ecology (31 articles total): high (+22 articles - green); medium (10-21 articles - yellow); low (<10 articles - red), none (white).

SOIL ECOLOGY **Overarching context categories**	**Practices**	**Common Context categories related to Practices**
Research Scope (31/31)	Collecting (31/31)	Sample (26/31)
		Variable/parameter (22/31)
		Instrument (19/31)
		Study location (28/31)
Data access (0/31)	Sampling (21/31)	Instrument (2/21)
		Study location (21/21)
	Reuse (4/31)	Data source (4/4)
		Citation (4/4)
	Processing (22/31)	Instrument (10/22)
		Citation (11/22)
	Analysis (31/31)	Variable/parameter (29/31)
		Modification (3/31)
		Software (18/31)
		Instrument (3/31)
		Citation (24/31)

Volcanology

The methods description from the set of Volcanology journal articles was the most robust for the *analysis* category compared with the other practice categories (see Table 9). Discussion of analytical processes centered on particular instrumentation used to conduct elemental, chemical, and whole-rock analysis. Specific techniques included X-ray Fluorescence (XRF) spectrometry, Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA–ICP–MS), and electron microprobe analysis. References to other labs are often detailed in the text when analysis is outsourced to employ complex equipment not owned in-house. Description from articles was also generally available for the MM context categories related to *analysis*, demonstrating the transferability of the practice-context category grouping from Soil Ecology to Volcanology.

In contrast to the consistency in description for *analysis*, the practices of *collecting, sampling, processing,* and *reuse* were not as well represented in the journal article content. Less

than half of the 15 total journal articles from the sample contained description related to these practices that could be included as methods metadata. The interviews with Volcanology researchers were able to supplement description for these categories to supply details needed to understand how data were generated for analysis. For *collecting*, the interviews revealed steps taken during a field campaign to locate and extract rock samples from a geologically significant location. These steps for *colleting* drawn from the interviews were useful for identifying gaps in methods information from journal articles, particularly when data production procedures were not explicitly described. The typical description provided in the Methods section of articles tended to focus on geological formation features of the field site, with minimal mention of how samples were actually collected for analysis.

The interviews also helped to confirm observations for *reuse* of data in Volcanology articles. Physical data such as thin section slides or rock samples are provided to colleagues based on personal requests (C1GEO1PI1_20100420; C1GEO1PD1_20100923), and it would therefore not be surprising that the analysis of these data would be published. However, reuse of data collected by researchers outside the laboratory group in Volcanology would not be considered a common practice, which accounts for the low count for *reuse* description in the set of journal articles.

Table 9: Availability of methods metadata from journal articles in Volcanology (15 articles total): high (+11 articles - green); medium (4-10 articles - yellow); low (<4 articles - red), none (white).

VOLCANOLOGY **Overarching context categories**	**Practices**	**Common Context categories related to Practices**
Research Scope (15/15)	Collecting (3/15)	Sample (2/3)
		Variable/parameter (3/3)
		Instrument (2/3)
		Study location (3/3)
Data access (3/15)	Sampling (5/15)	Instrument (3/5)
		Study location (2/5)
	Reuse (1/15)	Data source (1/1)
		Citation (1/1)
	Processing (5/15)	Instrument (2/5)
		Citation (5/5)
	Analysis (14/15)	Variable/parameter (14/14)
		Modification (0/14)
		Software (9/14)
		Instrument (5/14)
		Citation (7/14)

Stratigraphy

Similar to content from Volcanology journal articles, description for data *analysis* was most frequent in Stratigraphy articles (see Table 10). A prevalent type of analysis described was spectral analysis, which encompasses different techniques such as Lomb–Scargle (L–S) spectral analysis and the multi-taper method. The results from this analysis were often used to inform model design and development. As part of spectral analysis, the term "sampling" was often applied in regards to "sampling rate" or "sampling frequency" for estimations of noise in a time series. The same term was also used to describe sampling for physical samples such as cores from a geographic location. For this set of articles, *sampling* was only applied to procedures for physical samples, but the multiple uses of this term would need to be accounted for in broader efforts to extract methods metadata from journal articles.

Relative to Soil Ecology and Volcanology, there were more instances of data *reuse* described in journal article research from Stratigraphy. Nearly 50% of the 22 sampled Stratigraphy articles contained information related to use of existing data sources. Description for reuse included data source details and the enhancements made to the data to prepare them for analysis, as seen in the following example from a Stratigraphy research article:

> The second data subset includes DSDP, ODP, and IODP (Integrated Ocean Drilling Program) drill core records from other sites along the ROF (Fig. 1; Table DR1 in the Data Repository). The records extend back to ca. 1 Ma, but rarely include the shallowest sediments with ages up to ca. 100 ka, and have thus been complemented by data from gravity cores. These tephra layers are also dated using estimated sedimentation rates. (Kutterolf et al., 2013, p. 227-8)

In this example, multiple data sources contribute to one of the datasets for the study. The limitation of timescale values from the extracted data is acknowledged, which introduces the addition of data from gravity cores to better support the aggregated dataset for analysis. As confirmed by interviews with Stratigraphy researchers, the collection of physical samples from the Earth is immensely costly and time consuming, which makes using existing samples essential. Stratigraphy scientists also leverage personal networks to gain access to physical samples for analysis through project collaborations with funding for field campaigns. For methods metadata, descriptions of data reuse were consistent in detailing *data source* and *citation* information for these related context categories.

As mentioned above, the review of interviews with Stratigraphy researchers provided description of the circumstances surrounding data reuse, including existing data repositories

and the use of personal networks. Researchers also described the different kinds of software applied, particularly for time series data analyses. As with Volcanology and Soil Ecology, information related to *modification* was generally low or not provided in the Stratigraphy articles. Using interviews to generate information on *modification* would require a more targeted approach than the interviews conducted for this study.

Table 10: Availability of methods metadata from journal articles in Stratigraphy (22 articles total): high (+16 articles - green), medium (6-15 articles - yellow), low (<6 articles - red).

STRATIGRAPHY **Overarching context categories**	**Practices**	**Common context categories related to Practices**
Research Scope (22/22)	Collecting (13/22)	Sample (7/13)
		Variable/parameter (6/13)
		Instrument (4/13)
		Study location (12/13)
Data access (7/22)	Sampling (8/22)	Instrument (4/8)
		Study location (8/8)
	Reuse (10/22)	Data source (8/10)
		Citation (9/10)
	Processing (10/22)	Instrument (3/10)
		Citation (6/10)
	Analysis (19/22)	Variable/parameter (13/19)
		Modification (2/19)
		Software (14/19)
		Instrument (3/19)
		Citation (14/19)

Intra-case analysis: subdiscipline methods description

In developing the cases for methods metadata, recognition of "localized" practices was a key characteristic of subdiscipline research community practices. Examining how participants in the same research group described procedures in the generation and analysis of data contributed to understanding unique approaches to implementing methods procedures and the identification of descriptive components for methods metadata. Similarities in the techniques applied for sampling, data collecting, and processing across members from a single research group would suggest some evidence of localized practices. The identified techniques used across members could then be applied to metadata templates or integrated as part of a pre-populated form for making the metadata generation process less taxing for the data producer.

While the interviews were not expected to produce the same verbatim methods content from participants in the same research group, it was anticipated that common citations for particular procedures would at least be visible in journal articles. As one example of common citation use from Soil Ecology, participant C1SE1PI1 references "Raw, 1959" when describing the sampling process of earthworms from a site and the citation of the same process is also used by participant C1SE1GS1 who is a member of the same lab. This example, however, was a singular instance and there was generally very little overlap in references and common citation use in methods description from research articles published by participants from the same research group.

The diversity of citation use for methods exhibited in journal articles from a research group did not necessarily reveal shared localized practices. Even so, the methods description depicted by each participant in journal articles research groups provided multiple points of validation for methods metadata with the articles retrieved from the greater subdiscipline community. Those differences as well as similarities in the level of methods description availability from journal articles in a single subdiscipline are evident through the application of the MM Categories.

Methods Metadata Categories: Summary

The MM Categories are a set of terms to describe the processes of data production and analysis. The examination of journal articles and interviews established the initial foundation to identify methods metadata with the DPCVocab providing guidance on related research data practices. The categories that emerged represented not only the activities employed during the research process but also information on how these activities were enacted in a particular environmental setting. These two features of methods description are closely connected and distinguished as "practice" and "context" categories for methods metadata.

The analysis of journal articles, in particular, revealed patterns in the kinds of information consistently included when describing different processes of the methods. The availability of methods description from journal articles differed across the three subdisciplines, especially information on data collection procedures. There are evident connections between practice and content categories for methods description and the grouping of these categories can help make sense of the complex processes of data production and analysis. Distinctions in descriptive components attributed to some of the MM Categories for

guiding methods information provision were recognized for each subdiscipline. The localized practices of subdiscipline research groups for data production were not generally visible from journal articles and may be information better ascertained through interviews. With local practices more visible for methods description, portions of the metadata generation process could be tailored and even automated to accommodate the uniqueness procedures and techniques of the data producer and research group. The second phase of analysis discussed in Chapter 5 provides deeper examination of the identified MM Categories with metadata records and existing metadata schemes for scientific research data used in Earth Science data repositories.

CHAPTER 5: CORROBORATING METHODS METADATA CATEGORIES

The aim of Phase 2 was to corroborate the synthesized findings from MM category development with existing metadata records and associated schemes from data repositories. The examination of both the content of metadata records and formal metadata schemes provided a more holistic understanding of how the MM Categories could be adapted and applied to existing data standards and systems. Metadata schemes served as a baseline for identifying the aspects of the MM Categories that fit or could be added to current standards. The use of the metadata schemes proved to be key for determining gaps in methods description in both the MM Categories and repository records and served as a template for assessing what content journal articles and interviews can supply for the existing methods-related elements in a metadata schemes. In addition, content from metadata records illustrated current practice for how data production is documented and represented for existing datasets. Additional resources consulted to help verify the MM Categories included the National Environmental Methods Index (NEMI) and methods reference books from the respective domain sciences.

'Methods' coverage in data repository metadata schemes

In the pilot study detailed in Chapter 2, I examined if methods-related elements were included in existing metadata schemes for scientific data. As noted from that study, some metadata schemes contained methods-related elements for methods description even if there was not an element plainly named "methods." Some schemes were more explicit on what information about methods to provide while methods elements in other schemes had more ambiguous definitions for methods description. For the corroboration of MM Categories, Table 11 provides an overview of those metadata standards applied by the selected Earth Science data repositories together with elements likely to contain methods information based on element definition. The identified metadata elements are further discussed in the following section.

There was varied support for methods description from the Earth Science metadata schemes but overall, there was opportunity for information about methods to be included in a metadata record through the identified methods-related metadata. Of the examined schemes, the EML metadata scheme has the most extensive coverage for documenting methods with a specific "methods" module comprised of 16 different elements dedicated to description of

procedures implemented for data collection and processing.[22] The metadata standard for the EarthChem portal also has a designated section for methods description, "Method," encompassing several sub-elements for describing instruments and analytical procedures used.

There was less support for methods description using the PANGAEA standard. The description expectations for the "Method" consisted of the full name of the method and URL for a publication or a document archived with ePIC (electronic Publication Information Center).[23] Related to "Method" was "Event," which contained several sub-elements (e.g., date, longitude, latitude, etc.) adding context information to some of the processes employed during data generation. The DIF standard does not include a methods element, but information associated with methods could be included as part of the "Summary" element. Likewise, the definition for the Quality element — "information about the quality of the data or any quality assurance procedures followed in producing the data described in the metadata" (DIF Writers Guide, 2015)— suggested methods information might be presented. The DIF metadata scheme has also been mapped or cross-walked to the CSDGM and ISO19115 metadata standards meaning that retrieved records from GCMD can be viewed in these additional forms. The combination of DIF, CSDGM, and ISO19115 metadata elements were part of this analysis to encompass the breadth of areas where methods description may appear. Additional metadata elements for methods description from both CSDGM and ISO19115 included "Lineage" (information about events involved in dataset construction) and "Attribute_Accuracy_Report" (explanation of data quality and tests used).

The GSA repository metadata scheme was the only one without accommodation for methods description. There was generally minimal descriptive information for datasets and none of these elements could explicitly be used to capture methods description. The GSA data repository is closely connected with the society's publications database, which uses a much richer metadata structure for describing the publications. In this respect, datasets are not afforded the same level of support for access as compared with the scholarly research article.

[22] "eml-methods" module documentation, https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/./eml-methods.html

[23] ePIC, an open access publication repository hosted by the Alfred Wegener Institute for Polar and Marine Research, http://epic.awi.de/information.html

Table 11: Methods-related elements from data repository metadata schemes.

Repository: **Global Change Master Directory (GCMD)**	Metadata: **DIF (transforms to CSDGM, ISO19115)**
Methods-related elements and definitions: [DIF] "Summary" - A brief description of the data set along with the intended use of the data" including "scientific methodology or analytical tools." "Quality" - provide information about the quality of the data or any quality assurance procedures followed in producing the data described in the metadata. (DIF Writers Guide, 2015) [CSDGM/ISO19115] "Lineage" - information about the events, parameters, and source data, which constructed the data set, and information about the responsible parties. "Attribute_Accuracy_Report" - an explanation of the accuracy of the identification of the entities and assignments of values in the data set and a description of the tests used. (FGDC, 1998; ISO Data Quality, https://geo-ide.noaa.gov/wiki/index.php?title=ISO_Data_Quality)	
Repository: **PANGAEA (PANGAEA)**	Metadata: **repository standard**
Methods-related elements and definitions: "Method" table - definitions of analytical equipment, tools, or publications describing a specific method. (http://wiki.pangaea.de/wiki/Method) "Event" - defining the location for sampling or measurements (synonyms: site, station). If data are georeferenced, the event label and latitude/longitude are mandatory. (http://wiki.pangaea.de/wiki/Event)	
Repository: **Knowledge Network for Biocomplexity (KNB)**	Metadata: **EML**
Methods-related elements and definition: "eml-methods" module - Describes methods followed in dataset creation, including field collection, laboratory and processing steps, sampling methods and units, quality control procedures. Metadata elements include: dataSource, sampling, studyExtent, coverage, description, samplingDescription, spatialSamplingUnits, citation, referencedEntityId, qualityControl, protocol, instrumentation, software, subStep, ProcedureStepType, methodStep	
Repository: **EarthChem Portal (EarthChem)**	Metadata: **repository standard**
Methods-related elements (no definitions identified): From the data submission templates: "Method-specific Metadata" (bulk sample isotopic data), information about detection limit, total procedural blank, normalization, fractionation (if applicable). Additional fields can be added. For microprobe data: detection limit, total procedural blank, operation, and calibration. "Primary Analytical Metadata," information about the analytical procedure and accuracy and reproducibility including technique, laboratory, and reference sample information. From the "Method" section of the public metadata record: Method (Code, Name, Location, Provided by, Comment, Items measured); Precision (Item, Precision type, Minimum, Maximum); Standard sample measurement (Item, Sample name, Value, Stdev, Stdev type, Unit); Measured values have been normalized to 'normalization' (Item, Standard, value); Fractionation correction 'Isotopes' (Item, Fcorr Ratio, Standard, Value); Sampling (Field Program/Cruise, Date, Chief Scientist, Technique, Station).	
Repository: **Geological Society of America (GSA)**	Metadata: **repository standard**
Methods-related elements and definitions: (None identified; descriptive information includes Publication name, Year of publication, Authors, Title of article, and a unique identifier)	

Data repository metadata records analysis

The overall support for methods description in metadata schemes prompted the examination of how these methods-related elements were actually used. The content provided in a metadata record could help gauge the feasibility of extracting methods description from journal articles for these records. Metadata records were retrieved from each data repository and the content was examined for the methods-related elements defined in Table 11. The results of the metadata record analysis from all data repositories are detailed in Appendix E.

One of the main findings of this analysis was the direct attribution or use of journal articles in the metadata records for methods description. For some data repositories, the reference to a journal article was expected. Based on PANGAEA documentation for the Method element, a URL to a publication is part of the methods description; all but one of the URLs retrieved from the PANGAEA metadata records linked to journal articles. Figure 2 represents part of a PANGAEA metadata record retrieved for Volcanology. Of the two descriptions provided for Method, "ODP sample designation" resolved as a PDF file detailing the identifier used for cores, samples, and sections collected from the research site and "Prompt gamma neutron activation (PGNA Yonezawa et al. 1999)" resolved to a publication in the Journal of Radioanalytical and Nuclear Chemistry authored by Yonezawa and colleagues.

Parameter(s):

#	Name	Short Name	Unit	Principal Investigator	Method	Comment
1	Event label	Event				Metadata
2	Sample code/label	Sample code/label		Vils, Flurin		
3	DEPTH, sediment/rock	Depth	m			Geocode
4	Sample code/label 2	Label 2		Vils, Flurin	ODP sample designation	
5	Silicon dioxide	SiO2	%	Vils, Flurin	Prompt gamma neutron activation (PGNA, Yonezawa et al. 1999)	
6	Silicon dioxide, standard deviation	SiO2 std dev	±	Vils, Flurin	Prompt gamma neutron activation (PGNA, Yonezawa et al. 1999)	
7	Water in rock	H2O	%	Vils, Flurin	Prompt gamma neutron activation (PGNA, Yonezawa et al. 1999)	
8	Water in rock, standard deviation	H2O std dev	±	Vils, Flurin	Prompt gamma neutron activation (PGNA, Yonezawa et al. 1999)	
9	Hydrogen	H	%	Vils, Flurin	Prompt gamma neutron activation (PGNA, Yonezawa et al. 1999)	
10	Haplotype diversity, standard deviation	h std dev	±	Vils, Flurin	Prompt gamma neutron activation (PGNA, Yonezawa et al. 1999)	

Figure 2: Example of "Method" representation from PANGAEA for a Volcanology-related dataset; from Vils, F. et al. (2008): (Table 5) Whole rock chemistry of ODP Holes 207-1272A and 207-1274A. doi:10.1594/PANGAEA.783680.

Other instances of journal article reference were seen in methods description from the Summary and Quality elements from GCMD metadata records and the Description element from KNB metadata records. Some of the references accompany metadata text as sources for more information regarding a particular procedure but other times the metadata field just contained the journal article citation. The challenge with including just a citation as methods description is the lack of explanation for what information from the journal article is supposed

to apply for the respective element. More surprising was when description for the metadata element was taken directly from a journal article, as seen in this example from a Soil Ecology GCMD repository metadata record for the Quality element:

> "Taken from the referenced paper: Fieldwork was carried out on two separate occasions from the 3rd to the 19th of February 2004 and from the 27th to the 29th of February 2004…"[24] (ASAC_2397).

The statement of "taken from the referenced paper" denotes a journal article citation for "Greenslade, P., Stevens, M.I. and Edwards, R. (2007). Invasion of two exotic terrestrial flatworms to subantarctic Macquarie Island. Polar Biology 30: 961-967," which was information included as part of the metadata record. Delving into this referenced paper, the exact text used (i.e. "Fieldwork was carried out"…) to populate the Quality element for the metadata record is observed. Although these instances of journal article reference and content use only occurred in 11% of all retrieved metadata records, there was at least one record retrieved from each subdiscipline that contained this type of journal article information. Based on the subdiscipline observations in Chapter 4, there is the potential to extend the use of journal article content for methods description to these methods-related metadata elements.

In general, there was minimal methods information provided in metadata records. The use of methods-related elements for metadata description was also low across repository records. For instance, only information for four of the EML methods elements was consistently present in the retrieved metadata records. As would be surmised, those methods-related elements designated as "required" or "mandatory" in the metadata scheme were more frequently completed than optional, non-required elements. This was evident in GCMD records: Summary is a required element that had information completed for each record while information for the non-required Quality element had a much lower completion rate. The low use and completion of methods metadata elements are not surprising since scientists generally do not use metadata standards but opt for more localized practices in documenting data generation (Wallis, Rolando, & Borgman, 2013; Mayernik et al, 2012).

The potential application of journal article content for methods metadata may vary across repositories. Based on the metadata records with actual information completed for

[24] GCMD entry ID: ASAC_2397. Introduced invasive terrestrial invertebrates on Macquarie Island: studies on ecology, origins and control. Retrieved from http://gcmd.nasa.gov/KeywordSearch/Metadata.do?Portal=amd_au&KeywordPath=&EntryId=[AADC]ASAC_2397&MetadataView=Full&MetadataType=0&lbnode=mdlb2

methods elements, the contents were primarily in a narrative form similar to the writing style of a journal article. The exceptions to this were metadata records from the EarthChem and PANGAEA repositories. For PANGAEA, the element Event contains a number of sub-elements where numeric values are expected. The other sub-elements were text-based fields allowing brief descriptions, which contrasts with the narrative entries for methods-related metadata elements from other data repositories. The methods metadata for EarthChem also have a mix of numeric and text-based fields with similar expectations for brief descriptions. The description for the Technique element utilizes a controlled vocabulary maintained by EarthChem and is one of the few examples of a vocabulary specific to methods. The integration of this vocabulary with *analysis* from the MM Category, for instance, could help with locating the different techniques used for data analysis in journal article text. I discuss journal article use for methods metadata in the following section.

Subdiscipline observations: Metadata record generation

The generally low availability of methods information in metadata records inspired the examination of how methods description from journal articles could be leveraged to supplement the metadata generated for these records. Building on the analysis of Chapter 4, journal articles from some disciplines may be more conducive for supplying methods description in metadata records. I also consider how the metadata schemes can contribute to supporting subdiscipline differences in data production processes.

Each table accompanying the subdiscipline analysis contains the methods-related metadata elements visible in metadata records retrieved from data repositories. As noted in the previous section, not all methods-related metadata elements from a metadata scheme were regularly used for methods description. For each metadata element, I identified MM Categories representative of the kinds of methods information that may be gathered for completing a particular element. The MM Categories retain the color code assigned in Chapter 4 to show the level of methods description availability from subdiscipline journal articles in relation to the metadata elements. For some metadata elements, "(all)" is used to convey information from all MM Category groupings could potentially be included for methods description. In these cases, I consider how journal articles from that subdiscipline could supply methods information for that element. Table 15 provides a complete listing of MM Category alignment with methods-related elements from metadata schemes.

Soil Ecology

The journal articles from Soil Ecology were the most rich in methods description relative to Volcanology and Stratigraphy articles. The general availability of methods information from journal articles suggests methods-related metadata for PANGAEA, GCMD, and KNB have a high chance of being completed (see Table 12). In addition, the more general metadata elements such as Methods (PANGAEA) or Summary (GCMD) would also be fulfilled by the journal article content. The "methods module" from EML contains metadata specific to describing methods and the Soil Ecology journal articles were consistent in delivering methods description for data collecting, sampling, processing, and analysis, which yielded useful information for an EML metadata record. Additional EML methods elements such as Software and Citation match with the context-related MM Categories; based on Chapter 4 findings, Soil Ecology articles provided information for both Software, seen in relation to data analysis procedures, and Citation, as affiliated with processing data. The specific attention of EML to methods description provided a more concrete foundation for documenting data production processes than the elements from PANGAEA or DIF metadata schemes. It also seems feasible for journal articles from Soil Ecology to be used for a variety of methods-related metadata elements across the different schemes.

Table 12: Reconciling MM Categories with existing methods-related metadata elements for Soil Ecology research data. The (*) denotes a required element in the metadata scheme. The availability of methods description from journal articles is coded as high (green), medium (yellow), low (red), none (white).

SOIL ECOLOGY		Related MM Categories	
Repository (metadata standard)	**Methods-related metadata elements**	**Practice**	**Context**
PANGAEA (repository-specific)	**Method**	(all)	Variable/parameter
	Event	Sampling	Study location
		Collecting	Study location
GCMD (DIF/CSDGM/ ISO19115)	**Summary***	(all)	
	Quality	Processing	
	Attribute_Accuracy_Report	Collecting	Variable/parameter
		Analysis	
	Lineage	Collecting	
		Sampling	
		Processing	Citation
KNB (EML)	**Description***	Sampling	
	Instrumentation	(all)	Instrument
	Sampling Area And Frequency	Sampling	Instrument
			Study location
	Sampling Description	Sampling	Instrument
			Study location

Volcanology

There was not consistent methods description available from Volcanology journal articles for methods metadata elements (see Table 13). Information about how data were collected would be particularly difficult to ascertain from these articles although related context description such as information about where data collection took place (*study location*) could be identified. The general irregularity in methods description for practices suggests the general metadata elements from GCMD and PANGAEA may need to have information supplemented from other sources such as interviews.

In addition to PANGAEA and GCMD, metadata records were also retrieved from the EarthChem data repository. New metadata elements from this repository were introduced that were challenging to correlate with the MM Categories. Within the metadata records, information for Precision, Standard sample measurement, Normalization, and Fractionation correction (Isotopes) were numeric values which would generally not be accommodated by information provided by the MM Categories. In this respect, the set of MM Categories may not be the most appropriate indicator for methods description that could be contributed to these metadata elements requiring numeric values. The exception might be numeric values for longitude and latitude coordinates to describe *study location* but specific coordinates were not always included for the articles from each subdiscipline.

The EarthChem metadata scheme also contributed further guidance on methods information inclusion. The description in journal articles of electron microprobe use to determine element composition regularly detailed the voltage, beam current, and counting time of the device employed for this technique. The metadata element, Operation, contains sub-elements specific to voltage, beam current, and counting time, which is information we now know can be identified and extracted from journals articles. As previously mentioned, there is a controlled vocabulary in place for the EarthChem element, Technique. Evidence from Volcanology journal articles also points to the same abbreviations for techniques applied for analysis techniques, most notably X-ray Fluorescence (XRF) spectrometry and Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA–ICP–MS) analysis.

Table 13: Reconciling MM Categories with existing methods-related metadata elements for Volcanology research data. The (*) denotes a required element in the metadata scheme. The availability of methods description from journal articles is coded as high (green), medium (yellow), low (red), none (white).

VOLCANOLOGY		Related MM Categories	
Repository (metadata standard)	**Methods-related metadata elements**	**Practice**	**Context**
PANGAEA (repository-specific)	**Method**	(all)	Variable/parameter
	Event	Collecting	Study location
		Sampling	Study location
GCMD (DIF/CSDGM/ ISO19115)	**Summary***	(all)	
	Quality	Processing	
	Attribute_Accuracy_Report	Collecting	Variable/parameter
		Analysis	Variable/parameter
	Lineage	Collecting	
		Sampling	
		Processing	Citation
EarthChem (repository-specific)	**Sampling Technique***	Sampling	
	Method*	(all)	
	Precision	(none)	
	Standard sample measurement	(none)	
	Normalization	(none)	
	Fractionation correction (Isotopes)	(none)	

Stratigraphy

While methods description from Stratigraphy journal articles were relatively more consistent than Volcanology articles, only the description from the Attribute_Accuracy_Report element in metadata records from the GCMD would have methods information regular available from articles (see Table 14). Some of this contextual information represented in the Event element from PANGAEA records may potentially be gathered from journal articles. For example, the corresponding article to the dataset shown in Figure 3 offers the following details

that correspond with information for the Event sub-elements, "longitude" and "latitude," "Shatsky Rise, presently in the northwest Pacific Ocean (32°N, 158°E), is the only location with a known orbital lithology record of Paleocene age appropriate for investigating changes in atmospheric dust" (Woodard et al., 2011, p.6). While the longitude (158.505983) and latitude (32.651700) from the metadata record were more precise, similar coordinates were available from the journal article affiliated with the dataset. These geographic coordinates, as mentioned in the previous section, could be available description identified as part of the *study location* for collecting data but this detail was not consistent across journal articles. However, more general information for the study location was largely available from journal articles in the three subdisciplines and could potentially applied to the PANGAEA metadata.

Figure 3: Example of 'Event' metadata for a Stratigraphy-related data record from the PANGAEA repository; example from Woodard, SC et al. (2011): Sedimentary records of ODP Site 198-1209 on Shatsky Rise. doi:10.1594/PANGAEA.757701.

The metadata records retrieved from the GSA data repository did not yield any metadata elements related to methods description. It is unexpected that more robust metadata was not available for these data given the rich history of data provision with the GSA, which also publishes a number of professional journals for the geoscience community. The data repository originated in 1974 as the primary access location for datasets affiliated with a GSA journal article. It was surprising, therefore, that contemporary data repository practices have not been applied, especially where metadata is concerned. The datasets were linked to the GSA-published article with a unique identifier assigned by the repository but there was no indication that persistent links bridge these scholarly products. The publically accessible datasets were typically available in PDF, a challenging format for data reuse. Stratigraphy researchers discussed several tools and workarounds developed and used to digitally extract and analyze data points from these PDF files. As recognized by the interviewed scientists, the

journals of the GSA are a popular scholarly publishing venue. There is opportunity to build on the existing history of data sharing and adapt more robust practices for metadata provision.

Table 14: Reconciling MM Categories with existing methods-related metadata elements for Stratigraphy research data. The (*) denotes a required element in the metadata scheme. The availability of methods description from journal articles is coded as high (green), medium (yellow), low (red), none (white).

STRATIGRAPHY		Related MM Categories	
Repository (metadata standard)	**Methods-related metadata elements**	**Practice**	**Context**
PANGAEA (repository-specific)	**Method**	(all)	Variable/parameter
	Event	Collecting	Study location
		Sampling	Study location
GCMD (DIF/CSDGM/ ISO19115)	**Summary***	(all)	
	Quality	Processing	
	Attribute_Accuracy_Report	Collecting	Variable/parameter
		Analysis	Variable/parameter
	Lineage	Collecting	
		Sampling	
		Processing	Citation

Refining Methods Metadata Categories

The MM Categories were derived from empirical research of Earth Sciences subdisciplines to show the kinds of information needed for methods description. This set of terms was corroborated by methods-related elements from metadata schemes used by Earth Science data repositories. Based on the definitions for each MM Category, elements with similar definitions were identified from the metadata schemes (see Table 15). Each of the MM Categories seemed to fit with elements from the identified metadata schemes. I discuss the potential contributions of the categories to these existing metadata schemes as well as consider what gaps may still exist in this set of terms for methods metadata. The NEMI repository metadata scheme and descriptions of procedures from methods reference books added to the corroboration of MM Categories.

Table 15: Methods Metadata Categories alignment with existing methods-related elements, captured in parentheses, from identified metadata schemes.

MM Category	Metadata schemes with comparable elements
Collecting	ISO19115/CSDGM (Lineage); EML (Description)
Reuse	DIF (Summary); EML (Description)
Sampling	EML (Sampling Area & Frequency, Description); PANGAEA (Event); EarthChem (Sampling Technique)
Processing	ISO19115/CSDGM (Lineage); NEMI (Sample Handling, Sample Prep Methods); EarthChem (Normalization); EML (Description)
Analysis	DIF (Summary); EML (Description)
Research scope	DIF (Summary); NEMI (Scope and Application)
Study location	PANGAEA (Event); EarthChem (Method); EML (studyAreaDescription – not from Methods module)
Sample	DIF (Summary); NEMI (Media Name); EarthChem (Standard Sample Measurement); EML (dataSource)
Variable/parameter	CSDGM (Attribute_Accuracy_Report); PANGAEA (Method)
Data source	DIF (Summary); EML (dataSource)
Data access	NEMI (Source Citation)
Citation	NEMI (Method/Source Citation); EML (Citation, Protocol); ISO19115/CSDGM (Lineage)
Modification	DIF (Summary); EML (Description)
Instrument	EML (Instrumentation); NEMI (Instrumentation); EarthChem (Operation)
Software	EML (Software)

Types of methods metadata elements

The review of methods-related elements from metadata schemes showed support for methods description that complemented and expanded on the practice and context types initially identified for MM Categories. Five types of methods metadata elements emerged from the review of data repository metadata schemes and NEMI. The first two types, *practice* and *context*, consist of metadata elements that mirror existing MM Categories. The third element type, *combinatio*n, contains multiple sub-elements that accommodate methods description relating to both *practice* and *context*. The *combination* elements parallel the MM Category groupings. Metadata elements used for organizing methods description comprise the fourth

type of element. The *organization* elements were not always visible in the retrieved metadata records but are elements specified in some of the metadata schemes. The *general* elements constitute the final type of methods metadata elements and were not the most reliable for representing methods description.

The observed metadata elements types for methods description from existing metadata schemes are summarized below:

- *Practice* – single elements related to data production and analysis processes
- *Context* – single elements related to how a Practice was employed
- *Combination* – a grouping of elements that provide information related to the data sample/variable, related data production practices, and context information to situate these activities
- *Organization* – elements that contribute to connecting data and data production processes for methods description
- *General* – single elements that have the potential to contain methods information on data samples/variables, practices, and context

Practice and Context elements

MM Categories were based on metadata elements related to *practice* and *context* for methods description. The most prevalent and specific data production practice that emerges is "sampling" (identified from EML and EarthChem). In addition to practices related to sampling and the resulting data products are "sample handling" and "sample preparation methods" from NEMI. The presence of "sampling" and related activities from these metadata schemes in the Earth Sciences help to verify the inclusion of this term as a metadata category for methods metadata.

There was a greater frequency in the number of overlapping elements related to context information for methods description compared with practice elements. These elements broadly fit into the areas of "location," "instrument," "date," and "citation." While each metadata scheme has its own variation on the actual term used for the metadata element name(s), the inclusion of these themes across multiple metadata schemes advocates their importance for understanding the data production process. Information related to both "location" and "instrument" is included in the MM Categories. Location information was particularly well

described in the PANGAEA metadata with references not only to the study site and project campaign but also the option to include longitude and latitude coordinates. Instrument information was usually coupled with data practice activities; it was also an element reflected in the NEMI scheme as a set of pre-existing options for selection.

Another element related to context that was noted across the metadata schemes captures information about the time that data production processes occur, often represented as a "date." In relation to journal article content, there was more information related to the season or even specific months and the year when field campaigns occurred rather than an exact time and date. There may be embedded metadata from the actual dataset, which provides more granular detail about what temporal indicators to include for this element.

A final context element that frequently appears in metadata schemes is a reference or bibliographic citation for different aspects of the methods process. Citation information may be included for a specific technique, the data source used, or an associated publication providing more detailed descriptive information about the associated dataset. From the repository records analysis, the most common citations were references to journal articles. The prevalence of citation provision across these metadata schemes further enforces a best practice of documenting the provenance for a dataset in order to understand the context of its creation and use (Whyte & Wilson, 2010).

Combination elements

The *combination* elements bring together more concrete components for methods description that explicitly detail practices and context information identified in the research study. These elements also have the greatest level of alignment with the MM Categories not only in definition but also in function. Among the metadata schemes, EarthChem and CSDGM/ISO19115 (GCMD) contain representative elements that can be grouped as *combination* elements. Within EarthChem, the Method element not only comprises sub-elements for context information such as the location and "items measured" but also addresses specific processes such as "normalization" and "standard sample measurement." A comparable metadata element from CSDGM/ISO19115 is Lineage where information regarding specific process steps for a related data sample can be described and linked via a self-prescribed identifier. One of the sub-elements of Lineage is "citation" which is also a term represented in the MM Categories.

In the application of these elements, the metadata record content compliance analysis indicated that information was generally provided for the "Method" but not the other sub-elements from EarthChem records. Methods information was not available for Lineage from GCMD records. The EML metadata scheme contained similar elements for step-by-step process descriptions but these were generally not completed in the retrieved metadata records from KNB. Despite the minimal application of methods-related elements, it is clear that there is some support for methods description from existing metadata schemes.

Organization elements

Related to the *combination* elements are descriptive fields for organizing methods information. Examples of these *organization* elements for support categorizing and making methods information more manageable to locate and access include Method number/Identifier, Method descriptive name, as well as Method type/Subcategory from NEMI. The inclusion of an assigned identifier also appears in the EarthChem, PANGAEA, and CSDGM/ISO19115 schemes to group together information on the data production process with the associated data product. These types of elements could be potential additions to the MM Categories to help structure and capture the multiple groupings that are possible with these categories.

General elements

The fifth type of methods description element is considered "general" as there is the possibility for the inclusion of information related to data production and analysis processes but few concrete criteria to ensure this description is provided. The most prominent example of a *general* metadata element from the data repositories is "Summary" (DIF). Other elements such as "Method" (PANGAEA) concentrated on the name of the method, relying on external documentation to convey how the method was actually implemented. These types of elements can vary in terms of the amount of information provided that directly relates to the data sources and processes involved. The practice-related MM Categories were developed to accommodate the variation in procedures used by different researchers for producing data, which would be masked if this description were included in a general metadata element.

"Quality checking" as methods description

One potential addition to the MM Categories is a description of practices and indicators for checking the quality of data generated. The description of processes related to data quality is technically part of the definition for *processing* in the MM Categories but "data quality" appears as an individual metadata element in EML, DIF/CSDGM/ISOI9115, and NEMI metadata schemes. The available definitions for this element from metadata documentation converged on themes related to how quality of data is fulfilled during the research process based on implemented standards for quality checking and assurance. There were additional elements from NEMI that reveal some research area-specific descriptive factors such as "interferences" and "precision description notes" for water quality assessment. These additional quality elements provide a space for details on potential contaminates and proposed remedies that would influence how a method was implemented or the steps and standards undertaken and utilized in verifying the accuracy of data produced.

Identifying quality checking procedures directly from journal article content was not as straightforward compared with recognizing sampling and analytical processes. There generally was not a heading or obvious section dedicated to data quality practices within the structure of an article. Participant interviews did provide some indicators on what stage of the research process data quality checking may occur. From Soil Ecology participants, the statistical analysis of data was one stage of the research process where data quality control takes place:

> "If there is a number that looks completely out of the range, we will go back into the notebook and see if that was the correct number or not, like I said, that's part of the statistical analysis. So, we will do that review if all the data meets all the requirements for a statistical analysis and that's also kind of like what they call data checking or data assurance." (P1SS1PD1_20100716)

Additional quality control measures may be implemented during the collection of physical samples. For instance in Soil Ecology, concerns of cross-contamination of samples were addressed during the collection phase by cleaning the collection instruments beforehand and verifying that the same instruments were used for samples extracted from a particular plot (P1SE1GS1_20100916a). Similar precautions were taken in handling collected samples—"the processing protocol should be the same for all samples, and field and laboratory replicates should be randomized before processing to minimize systematic errors (Boone, Grigal, Sollins, Ahrens, & Armstrong, 1999, p.11). From journal articles, the use of replicates was typically discussed as part of the sampling and experimental design (i.e. five replicates for each

treatment in a split-plot design) but was not necessarily recognizable as an indicator of quality control. Other quality control considerations from domain science method textbooks for Soil Ecology convey the importance of using reference materials, blanks, and spiked samples in the analytical process to ensure accurate calibration of laboratory instruments, protocols, and standards (local and certified) for measuring soil properties. For example, the use of reference soils from NIST, which have certified values for chemical properties regardless of method applied, was not observed in journal article content from the study sample. This observation corresponds with method textbooks for Soil Ecology that state localized standards and practices are common for quality control purposes (Boone, Grigal, Sollins, Ahrens, & Armstrong, 1999), which may account for the absence of reference soil description in articles.

For Volcanology and Stratigraphy research, identification of quality checking practices and indicators from journal articles and interviews was limited. One Volcanology researcher explained that checking for data quality might be trigged by the presence of an outlier when reviewing processed data (C1GEO1PD1_20100923). Other instances of quality discussion appear in methods texts as guidance on the proper care and calibration of instruments and equipment used in the field to make certain of sound data collection (Coe, 2011; Knödel, Lange, & Voigt, 2007). In a similar vein, there is an understanding that the quality of data is relative to the instruments at-hand for a field campaign (Low, 1957). As a Volcanology scientist described, the use of compasses, laser range finders, and human height as an elevation marker are some of the devices and techniques used for orienting to a field location and determining areas for sampling (C1GEO1GS1_20100604). The MM Category of *instrument* could be an important element to include for data quality description for Volcanology metadata given the significance of instruments and devices for establishing quality of data.

An indicator of data quality checking described by Stratigraphy participants was the inclusion of journal publications as part of the data documentation (C1GEO3PI1_20100924). Method textbook information for Stratigraphy research also confirms the importance of robust documentation especially when utilizing samples as reference points. For example, a well-documented reference well log for a borehole core would include lower and upper depths and maximum thickness of unit drilled, physical characteristics of the core, formation boundaries, lateral extent, and age (Rey & Galeotti, 2008). Some of these criteria are reflected in PANGAEA

metadata elements, such as Maximum and Minimum depth,[25] but these are not directly connected with the PANGAEA Method element.

Although "quality" description was not explicitly included as a heading in journal articles, there were indicators from article content of quality measures and practices descriptions. Identifying the processes or indicators related to data quality would likely require domain science expertise for review and verification. Even with minimal data quality description identified in journal article content, it does not mean researchers are not employing quality control processes. Other sources, such as a scientist's laboratory notebook, may contain greater detail on quality practices suitable for metadata description.

Methods description organization

The examination of metadata records and schemes for methods description also revealed elements for the structural representation of the methods process. Both CSDGM/ISO19115 and EML contain elements that allow for step-by-step description of methods through the inclusion of "methodstep" (EML) and "ProcessStep" (CSDGM/ISO19115). The use of these elements can bring greater clarity to the overarching processes of data production and analysis, yet the examined metadata records showed that these elements were not commonly applied in practice. Other organizational considerations for representing methods description were seen in PANGAEA and EarthChem records. The tabular format for description can add organizational structure to maintain the relationship between the name of a data parameter or variable and each methods process undertaken. This sequential and sometimes brief description format can also be easier to follow by data producers and users.

Figure 4 displays two representations of the same methods information for a Soil Ecology study based on journal article content. The top section (A) adapts the EML scheme for methods description and depicts each step of the data production process with information regarding the software or instrumentation used as well as related citations to referenced procedures. From the metadata record example (A), the element fields were completed using direct extractions from the journal article and generally follow the narrative of the article's Methods section. The application of "steps" helped to delineate the sequence of procedures

[25] Minimum and maximum depth information described in PANGAEA documentation: "Coverage" (see Fig. 6) http://wiki.pangaea.de/wiki/Data_set#Coverage_tab_.28Fig._6.29 and "Age" http://wiki.pangaea.de/wiki/Age

deployed in the study from an otherwise large block of text. The overlap between the MM Categories and EML methods elements is evident with methods description of *software*, *citation*, and *sampling*. Some of the practice-related MM Categories such as *processing* could also be applied for "Step 3" to further explicate the techniques utilized.

The bottom section (B) relays the same methods information and uses the PANGAEA table as a format guide. The identified data parameters are based on study results presented in the paper along with inferences made from available text. Ideally, an associated dataset would provide a more definitive listing of those parameters. Similar to PANGAEA, detailed accounts of how a procedure occurred would be part of a separate interface (i.e. link to separate documentation) and not part of the main methods metadata record. The table format also integrated some of the practice-related MM Categories to map out the research processes in relation to the data generated. The challenge for utilizing either format is in identifying and gathering the methods information for metadata generation. The narrative format makes greater use of the traditional style of journal article content to generate metadata; the tabular format would take longer to generate but would result in a more accessible format to review and be more amenable to automated processing for metadata generation.

A) narrative format for methods metadata

Methods

Step 1	
Description	We conducted a microcosm experiment under greenhouse conditions to assess the impacts of plants (Brachiaria decumbens) and earthworms (Pontoscolex corethrurus) on soil structure and C stabilization. Aggregate stability was assessed by wet-sieving. Large macroaggregates (>2 mm) were also visually separated according to origin (e.g., earthworms, roots) and then further fractionated into particle size fractions to assess aggregate composition and C distribution. This experiment was conducted using soil microcosms (17.5 cm dia. × 17 cm tall, fitted 1 mm nylon mesh to prevent earthworm escape) under greenhouse conditions at the International Center for Tropical Agriculture (CIAT) near Cali, Colombia. Soil, collected at the CIAT campus (3°30′ N, 76°20′ W), is classified as a Cumulic Haplustoll (Howeler 1986) and has a silty loam texture (23 % sand, 53 % silt, and 23 % clay), total C content of 13.5 gC kg−1 soil and a pH of 7.2. The soil was air-dried and passed through an industrial mill and a 2 mm sieve to ensure the complete destruction of large macroaggregates and removal of stones and large organic residues. The soil was then mixed with sand, to ensure adequate drainage, at an equal ratio (1:1) of sand to soil, and then 2.5 kg of this mixture was added to each microcosm and packed down by hand to a uniform level to ensure roughly equivalent bulk density across microcosms. The microcosms were thoroughly watered from below (via capillary action) to ensure even wetting of the entire soil profile and allow for settling of the soils and mixture before treatment establishment.
Step 2	
Description	One half of the soil was used for aggregate fractionation by wet-sieving and determination of aggregate stability, while the other half was separated visually into morphological structures according to their origin (see details below). The soil from both halves was carefully inspected to recover coarse roots and earthworms. Earthworms were washed and left to void their guts for 48 h and then weighed to determine weight change over the course of the experiment. The coarse roots were washed to remove soil particles and dried, along with aboveground biomass, in an oven at 60°C and then weighed for determination of above and belowground biomass in each treatment. Aboveground biomass was ground and analyzed for total N and P content (Jones et al. 1991).
Step 3	
Description	Upon harvest, the moist soil designated for wetsieving was passed through an 8 mm sieve by gently breaking aggregates along natural planes of weakness and then air-dried. Aggregate fractionation by wetsieving was conducted following methods adapted from Elliott (1986) to obtain four aggregate size fractions: 1) large macroaggregates (>2000 µm), 2), small macroaggregates (250–2000 µm), 3) microaggregates (53–250 µm) and 4) silt and clay (<53 µm). In short, 80 g of air-dried soil was submerged in deionized water on top of a 2 mm sieve.After slaking for 5 min, the sieve was moved up and down with an undulating motion 50 times in 2 min. Soil remaining on the sieve was then rinsed into a pre-weighed aluminum pan, while soil passing through the 2 mm sieve was transferred to a 250 µm sieve and the process repeated again with both the 250 µm and 53 µm sieve, until the four fractions were each in their respective pre-weighed pans. The contents of each pan were dried at 60°C and then weighed to determine the proportion of soil in each size fraction.
method citation	[wetsieving] Elliott ET (1986) Aggregate structure and carbon, nitrogen, and phosphorus in native and cultivated soils. Soil Sci Soc Am J 50:627–633
Step 4	
Description	Soil was separated by visual inspection according to Velasquez et al. (2007) to yield three classes of large macroaggregates of different origin: 1) earthworm casts distributed throughout the soil profile (CAST), 2) rhizosphere associated aggregates strongly adhering to plant roots (RHIZ), and 3) aggregates formed by physiochemical interactions and microbial processes (PHYS), as well as non-aggregated soil and aggregates smaller than 2000 µm(NON). Upon separation each fraction was airdried and weighed in order to determine the relative proportion of the whole soil represented by each fraction. A subsample of each fraction was ground and dried for analysis of total C using a modified Walkley and Black method (Rabenhorst 1988). Large macroaggregates of different origins (CAST, RHIZ and PHYS) were further separated using mechanical disturbance and wet-sieving (see Six et al. 2000) to obtain three size fractions: 1) coarse particulate organic matter (>250 µm; cPOM), 2) microaggregates within the macroaggregates (53–250 µm; mM), and 3) silt and clay within the macroaggregates (Msc).
method citation	visual inspection according to Velasquez et al. (2007)
Step 5	
Description	The influence of plants and earthworms on soil aggregation (as determined by wet-sieving), as well as C storage and the composition of the morphological aggregate fractions was examined using ANOVA, with the earthworms, plants, and the earthworm × plant interaction considered the main effects. All data were checked to meet the assumptions of ANOVA and natural-log transformations were applied as necessary to satisfy these assumptions. Comparisons of means between fractions and treatments were performed using Tukey's honestly significant difference.
Software	Analyses were conducted using JMP 9.0 statistical software (SAS Institute 2010).

Sampling

Sampling Step 1	
Sampling Area & Frequency	soil microcosms (17.5 cm dia. × 17 cm tall, fitted 1 mm nylon mesh to prevent earthworm escape) under greenhouse conditions at the International Center for Tropical Agriculture (CIAT) near Cali, Colombia.
Sampling Description	At harvest, plants were clipped at the soil surface and soil in the microcosms was destructively sampled by dividing the microcosms into two equal halves by carefully parting the soil down the middle with a knife and using scissors to cut large roots.

B) tabular format for methods metadata

Summary

Contributions of soil macrofauna and plants to soil aggregation and C storage, relatively little work has been conducted to understand how these two fundamental soil components interact to influence SOM dynamics in agroecosystems. This experiment was conducted using soil microcosms (17.5 cm dia. × 17 cm tall, fitted 1 mm nylon mesh to prevent earthworm escape) under greenhouse conditions at the International Center for Tropical Agriculture (CIAT) near Cali, Colombia. Soil, collected at the CIAT campus (3°30′ N, 76°20′ W), is classified as a Cumulic Haplustoll (Howeler 1986) and has a silty loam texture (23 % sand, 53 % silt, and 23 % clay), total C content of 13.5 gC kg−1 soil and a pH of 7.2.
* indicates additional details are available in the journal article

Sample Source

Soil microcosm - (4) treatments x (5) replicates)

Data Parameters	Gathering	Processing	Statistical Analysis	Cited References
C storage	destructive sampling of soil microcosm*	from soil fractions, used modified Walkley and Black method (Rabenhorst 1988)	ANOVA, earthworms, plants, and the earthworm × plant interaction considered the main effects; JMP 9.0 statistical software (SAS)	Rabenhorst MC (1988) Determination of organic carbon and carbonate carbon in calcareous soils using dry combustion. Soil Sci Soc Am J 52:965–969
soil aggregation	destructive sampling of soil microcosm*	wet sieving (Elliott 1986)*; mean weight diameter calculations	ANOVA, earthworms, plants, and the earthworm × plant interaction considered the main effects; JMP 9.0 statistical software (SAS)	Elliott ET (1986) Aggregate structure and carbon, nitrogen, and phosphorus in native and cultivated soils. Soil Sci Soc Am J 50:627–633
morphological aggregate fractions	destructive sampling of soil microcosm*	1) visual separation (Valesquez et al, 2007); 2) composition and distribution of C: mechanical distribution & wet seiving (Six et al, 2000)*, dry combustion (Rabenhorst, 1988)	ANOVA, earthworms, plants, and the earthworm × plant interaction considered the main effects; JMP 9.0 statistical software (SAS)	Velasquez E, et al. (2007). This ped is my ped: visual separation and NIRS spectra allow determination of the origins of soil macroaggregates. Pedobiologia 51:75–87; Rabenhorst, MC. (1988). Determination of organic carbon and carbonate carbon in calcareous soils using dry combustion. Soil Sci Soc Am J 52:965–969
earthworm- weight change	destructive sampling of soil microcosm*	Earthworms were washed and left to void their guts for 48 h and then weighed		
aboveground/belowground biomass	destructive sampling of soil microcosm*	coarse roots were washed to remove soil particles and dried, along with aboveground biomass, in an oven at 60°C and then weighed		
Total N and P content	destructive sampling of soil microcosm*	Aboveground biomass was ground and analyzed (Jones et al., 1991)		Jones JB, Wolf B, Mills HA (1991) Plant analysis handbook: a practical sampling, preparation, analysis and interpretation guide. Micro–macro Publishing, Athens

Figure 4: Examples of narrative (A) and tabular (B) formats for methods description; methods text from Fonte, S. J., Quintero, D. C., Velásquez, E., & Lavelle, P. (2012). Interactive effects of plants and earthworms on the physical stabilization of soil organic matter in aggregates. Plant and Soil, 359(1-2), 205–214.

Methods Metadata Categories corroboration: Summary

The MM Categories were verified with several sources including existing metadata schemes from data repositories and established scientific domain practices for methods. The analysis of existing metadata records helped to further corroborate journal article content as a source for generating methods metadata. Across the three subdisciplines, references to and direct content from associated journal articles were observed in metadata records. The various metadata schemes applied by the data repositories also revealed a range of support for methods description. On the whole, the EML utilized by the KNB showed the greatest alignment with the MM Categories, corroborating the empirically derived MM Categories since EML is one of the more comprehensive schemes for coverage of methods. This alignment provides evidence for journal article content use, especially from Soil Ecology, to fulfill existing metadata schemes for methods-related description. In this respect, the many EML elements for methods description that were not observed in the retrieved metadata records could potentially be populated with journal article content.

Methods-related elements from each metadata scheme were not necessarily readily identified or used in metadata record generation. In some cases, the same term appeared in multiple metadata schemes but was either not connected with methods description or had broader use (i.e. instrument). The number of metadata records with fairly complete methods description based on available elements was generally low. There did not appear to be any significant differences in methods information for records retrieved from the more data-specific repositories for Volcanology (EarthChem) and Stratigraphy (GSA) compared with the broader Earth Science repositories of PANGAEA and GCMD. However, methods description was relatively more complete from Soil Ecology metadata records (EML) than in records from those broader Earth Science repositories.

Methods Metadata gaps and additions

The examination of existing metadata schemes from data repositories and the addition of NEMI for methods description revealed both areas the MM Categories were lacking or could contribute to the current schemes. The contributions of the MM Categories center on making visible the different stages of the research process and those associated practices for data production and analysis. Sampling was a prevalent practice for field research represented in metadata schemes, but the related activities of analysis and processing of data were not

explicitly represented in the schemes. These practices have associated instruments and procedures that may be overlooked for metadata provision in other schemes.

The primary area not represented in MM Categories was description for data quality checking practices and assurance. The identification of such practices from journal article content proved to be challenging without domain expertise in the standards and references that would indicate practices to ensure the quality of data. As a methods-related element, data quality activities may already be embedded in existing protocols and would require domain knowledge to locate and extrapolate these actions for metadata description. The identification of additional types of methods metadata from the schemes showed the issues of having generic definitions for *general* elements but also recognized the need for *organization* and *combination* elements to handle complex MM Category groupings.

Methods Metadata Organization

Another consideration with metadata in general and with the MM Categories in particular is how methods metadata can be represented. The retrieved records portray methods description in a value-based tabular format or in a narrative style. The advantage of the narrative style is the flexibility in directly extracting and referencing journal article content for metadata generation. This process is made more tractable as the majority of methods description is located under a designated "Methods" section, especially for Soil Ecology articles. The disadvantage of this representation is the influx of information that can be included making the completed record cumbersome to follow and read; it may be the case that reviewing the actual journal article would be more helpful than in the record form.

Utilizing the value-based approach offers a more abstracted view of the methods process with description at a minimum but augmented through external links. Integrating a tabular structure for methods representation also brings together information on what and how data were produced that is more visually engaging to follow. This highly structured approach may better lend itself to use of controlled vocabularies and automated processing. Developing these types of records, however, may initially be more time intensive to complete, as many of the processes and descriptive components need to be interpreted and re-configured for record completion.

CHAPTER 6: DISCUSSION & CONCLUSION

This chapter presents a discussion of major findings from the preceding two chapters guided by the two research questions for this study:

a) What methods metadata relating to how data are gathered, processed, and analyzed for research can be derived from qualitative interviews or from research journal articles?
b) How does methods metadata differ across subdisciplines in Earth Science?

I discuss the implications of these findings and conclude with future directions.

Summary of findings

Methods metadata convey description of procedures employed for producing and analyzing data. The MM Categories, empirically situated in analysis of journal articles and interviews, provide an initial set of terms for identifying both practice and context information critical for methods description. The multiple relationships that comprise methods (seen in the category groupings in Chapter 4) for driving new scholarship shows that there is much more to methods description from journal articles than just the name of a method. As suggested by the *organization* and *combination* metadata elements identified in Chapter 5, there is also a structure to how methods description can be represented to show each procedure used in relation to the data generated.

MM Categories as "reference"

The findings of this research highlight the MM Categories as an example of a "reference" in Latour's (1999) concept of *circulating reference* for the systematic formation of knowledge. The development of the MM Categories from journal articles and interviews represents the series of transformations characteristic of a reference. This progression of changes is evident in the iterative analysis of the articles and interviews, which involved with multiple rounds of coding to develop the initial set of MM Categories for methods description. A reference may result from the synthesis or "reduction" of information but also be used to emphasize or "amplify" unique attributes. The basic function of the MM Categories is to assist with determining methods description from published journal articles. The different methods described in articles would be reduced by the MM Categories to show the practices involved in data production. Identified methods metadata would be further amplified by the MM Categories through the mapping of categories with existing metadata elements for more robust

generation of metadata. The attention to practices is a notable feature of these categories as current metadata descriptions of methods often overlook how the data are handled, processed, and analyzed. Given the alignment of MM Categories with metadata schemes, this set of categories also makes visible the commonalities between repository-specific and discipline-centered metadata schemes for scientific data. The potential exists therefore to make methods description more standard in metadata schemes for datasets to facilitate future use. The circulating aspect of this reference can be summarized by "methods and scientific knowledge thus progress in parallel, with each area of knowledge contributing to the other" (National Academy of Science, 1995, p. 4). The set of MM Categories will continue to inform methods metadata as new methods are created and evolve.

Metadata support for "sampling" and "reuse" description

In addition to description for how data are gathered, processed, and analyzed, the sampling and reuse of data were distinctive practices conveyed in methods description. Information for sampling was accommodated in multiple metadata schemes and available from journal article content in all three subdisciplines. Articles from Soil Ecology research, in particular, included a section specific to description of the sampling procedure conducted in the field. The attention to sampling confirms findings on the importance of documenting sampling procedures to support reuse of data from geobiologists engaged in site-based research (Thomer et al., 2014). This report on the geobiology community also contributed specific information for describing sampling procedures, which includes the quantity and replication of a sample along with the tools used. Based on the descriptive components associated with the *sampling* MM Category, information for replication of a sample and tools was consistent with Soil Ecology methods description from journal articles while quantity of a sample was more commonly described in Volcanology and Stratigraphy articles. The sampling description from journal articles can serve as a starting point to ensuring information on sampling procedures are made available with a dataset.

Data reuse methods described in journal articles provided insight into descriptive information common for understanding what data sources are used and the state of practice for how these data sources are actually retrieved, processed, and prepared for new analysis. Interestingly, there was some indication of subdiscipline differences for reuse methods description between studies using a meta-analysis approach and those reuse studies using

other techniques for analysis. The meta-analysis methods description from a Stratigraphy study contained no concrete details of how data sources were selected and retrieved, which contrasts in the high level of detail for methods provided in other types of reuse studies from this subdiscipline sample. However, meta-analysis research described in Soil Ecology journal articles had more thorough information on how data were searched for (i.e. use of keywords) and selected from existing repositories, what parameters were extracted from these data sources for analysis, and limitations of the aggregated data. The list of retrieved data sources are usually provided as a supplement to the study or included as a table or figure within the structure of the journal article.

Based on these reuse descriptions, the processing of data sources could be inferred from the discussion of data limitations and the steps taken to normalize and accommodate for differences in data representation from the diverse sources. The discussion of data limitations also seems to resemble the kind of information that would be provided for "data quality" metadata elements with emphasis on the approaches used to overcome the shortcomings of the data. Ensuring the quality of research synthesis relies on documenting procedures for dealing with missing data, and that methods used for extracting data are "justifiable, clearly documented, and repeatable"; providing such information for meta-analysis research especially within Earth Science research is still a work in progress with no formal practices established for describing methods in these instances of data reuse (Koricheva et al., 2013).

Interview contributions: methods standards and local practices

The interviews with subdiscipline researchers were an essential source for confirming observations of methods description from journal articles. Portions of the interviews helped to address gaps in description or enhance understanding of data production discussion, as seen with Volcanology and Stratigraphy articles. They were also informative sources for designating established or accepted techniques for different stages of data production. For instance, the use of mustard liquid and electro-shock techniques for extracting earthworm samples described in journal articles was confirmed as a standard approach in a Soil Ecologist interview for collecting this biological species from the soil environment.

One challenge in this study that may be addressed by interviews with data producers was assessing differences for methods metadata within a single Earth Science subdiscipline. The use of citations to track similarities in methods description from journal articles of

members from the same research group did not yield conclusive findings on the localized practices of a laboratory. It is anticipated from the literature (e.g. Wallis et al., 2013) that scientists in long tail science areas would utilize locally developed practices for documenting data generation but determining what practices were local or community-based was not feasible from methods description in journal articles. Understanding differences in methods metadata within a single research group let alone multiple research groups in the same subdiscipline would benefit from interviewing time. This finding is perhaps not surprising as qualitative interviewing was the approach used by Swan and Brown (2008) and Cragin et al. (2010) to study discipline differences for data practices.

Subdiscipline differences in methods metadata

The primary difference between the Earth Science subdisciplines was the availability of methods description from journal articles for methods metadata generation. Soil Ecology journal articles provided the most robust coverage of methods description that span practices of gathering and collecting data to data processing procedures and finally the techniques and approaches used to analyze data. The most consistent methods description from Volcanology and Stratigraphy articles was information relating to the analysis of data. While scientists in these three subdisciplines conduct field-based research and use data gathered in the field, there was not necessarily description of these processes in the journal articles from each subdiscipline.

The methods description differences were further enhanced by elements in metadata schemes for data repositories. The EarthChem metadata, in particular, illuminated both standard techniques deployed for data analysis through a controlled vocabulary and the description needed for instruments used in microprobe analysis; information on these techniques and instruments was contained in Volcanology journal articles could therefore further contribute to methods metadata for EarthChem data. It may be possible to extend these metadata elements to Stratigraphy or Soil Ecology metadata if similar techniques or instruments are utilized for research. Differences in methods metadata were also noted in descriptive components associated with data production practices. The guidance on information to include for methods description was specific to the subdiscipline based on common themes from journal articles and documentation practices and expectations for data sharing and reuse in interviews. The metadata scheme in this case may be a more accessible

source for understanding potential differences for methods description at the subdiscipline level. In general, the differences in practices reflected in methods description at the subdiscipline level confirm Cragin et al.'s (2010) findings on the level of analysis to understand data practices.

Implications for research approach and design

Overall, the combination of semi-structured interviewing and content analysis of journal articles can be an effective approach in the identification of methods metadata. Figure 5 depicts the relationship between these two approaches for gathering methods description for methods metadata in each subdiscipline. Based on the first phase of the study (see Chapter 4), the use of journal articles as a reliable source for methods metadata differed across the Earth Science subdisciplines. The designation of "high" or at least 70% of articles containing methods information was most prominent for Soil Ecology articles. Analysis of journal articles would therefore be a more effective technique for obtaining methods metadata with less time needed to interact directly with data producers from Soil Ecology. Any interviews conducted could be used to confirm the identified methods description from articles.

The strategy for identifying Volcanology methods metadata relies more on interviewing data producers about the methods used and less on journal article analysis since coverage of methods was not as consistent or standard. Relative to Soil Ecology and Stratigraphy, there were a greater number of journal articles for Volcanology that contained less than 30% of methods information. Interviewing data producers would likely be of greater assistance for procuring information on the data production and analysis procedures than depending solely on journal article analysis. Gathering methods metadata from Stratigraphy research depends equally on analysis of journal articles and interviewing for determining methods information. The "medium" level of availability characteristic of the Stratigraphy journal articles indicated that methods description from journal articles was not consistent. For example, methods information about *processing* or *collecting* data was available in 50-60% of journal articles necessitating interviews to supplement the information likely to be missing from Stratigraphy articles.

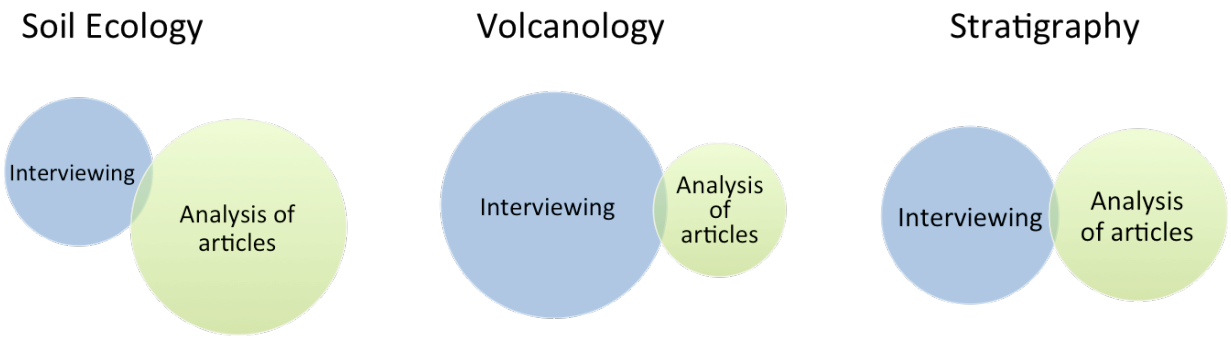

Figure 5: Combined approaches for attaining methods description for Soil Ecology, Volcanology, and Stratigraphy research data.

For information professionals tasked with curating research datasets, a journal article from the researcher with methods description could provide the foundation for creating the metadata and help make the best use of the time when directly speaking with the data producer. The expansion of this methods metadata study to other scientific research areas would require an understanding of the publications most likely to contain methods description. The most accessible and robust source for methods description was journal articles for the three subdiscipline cases but this source might be different in other communities of research. With the use of journal articles for this study, developing the sample set to include articles from both interview participants but also the broader research community proved to be an effective strategy for understanding variation in the types of studies conducted and range of techniques applied. More importantly, the overlaps observed for techniques and procedures across the sample of journal articles for a subdiscipline verified the sampling approach, which could be adapted for future studies.

Implications for metadata generation

It is clear that there is some support for methods description from existing metadata schemes, although at present few of these elements are being leveraged. Identifying methods metadata from journal articles is an initial step to populating a metadata record for use in an Earth Science data repository. While the availability of methods description varied across Earth Science subdisciplines, the differences in the physical structure of articles from a variety of journals and the vocabulary used for methods description are challenges to contend with for automating methods metadata generation from these articles. Similar to methods description

availability, subdiscipline differences also exist for the potential automation of generating methods metadata.

The ease of identifying and extracting relevant methods description has an influence on the application of automated approaches for metadata generation. Journals from each of the subdisciplines have different physical structures. The articles from Soil Ecology show the greatest consistency across journals for having a standard section "Methods and Materials" concentrated on discussion of research methods. Subheadings within this section provide additional structure in identifying the different steps of the methods process enacted during the research study. A subheading that regularly appeared in these articles was "Site description" which aligns with methods metadata for *study location*. In general, these subheadings were not uniform across articles but tailored to descriptions of the process. For instance, one article used an all-encompassing heading "Plant harvest and analyses" to convey information on processes for the collection of physical samples and the procedures used for analysis of chemical composition and microbial activity. Another article utilized individual subheadings for each analytical process, such as "Soil chemical analysis" and "Microbial community analyses." While the article structure tended to follow a sequential order (e.g., site description, data gathering, data processing, data analysis), the variations in subheadings would need to be accommodated if automation processes are adapted.

In the articles from the Stratigraphy sample, there was varied use of headings to indicate a content section specific to methods description. A comparable heading to "Methods and Materials" in Soil Ecology journals was "Data and methods" in some Stratigraphy articles. This section typically brought together information about the samples used for stratigraphic analysis with the processes for tuning data to explain different models of the Earth's rotation. Variations of "Methods" and "Methodology" for section headings were also visible and accounted for about half of the articles in the subdiscipline sample. In those articles without an apparent Methods section, alternative headings used denoted description of methods relating to "analysis" such as "spectral analysis" or "frequency analysis." The emphasis on a particular research activity such as analysis offers some indication of the presence of methods procedures being employed. As a complement to the "site description" subheading in Soil Ecology articles, a number of Stratigraphy articles had a "geological setting" section discussing the history and significance of the geological site being studied. The information on geological setting often preceded description of methods (if available) and unlike Soil Ecology articles, was not part of

the methods portion of the paper. It is not clear, therefore, if "geological setting" information would have the same function as "site description" in situating data production processes.

The journal articles produced by Volcanology researchers demonstrated limited use of "Methods" as a journal content section heading. The headings presented to signal methods content were more descriptive of the study content. Such section headings as "Quantitative textural measurements" or "Description of samples, data collection and analytical procedures" are examples of these descriptive representations indicating methods description without the explicit use of the term "method." These variations in headings seem to be specific to the data producer rather than standard headings used in research article publication. A few Volcanology articles contained a "geological settings" section, similar to the content section present in Stratigraphy articles. In the respect, the common terminology use shows greater similarity between Stratigraphy and Volcanology research publications than with content described in Soil Ecology publications.

The set of MM Categories can be a basis for developing a vocabulary that builds on the unique terminology used in journal articles for methods description. To a certain extent, the automated metadata generation may be more feasible for Soil Ecology journal articles than with articles from Stratigraphy or Volcanology given the initial ease to locate methods description. It may be prudent to start with articles from a single journal in piloting this approach for easing metadata frictions.

Implications for data publishing

The findings from this research raise considerations for how methods metadata from journal articles may be used in developing and potentially populating a data paper. In essence, methods metadata would be reused for data paper generation. Discussed in Chapter 2 as a growing venue for sharing data, the papers published in data journals focus on description of dataset generation. As stated in the objectives of the Geoscience Data Journal, a data paper "allows the reader to understand the when, how and why data was collected and what the data is."[26] The MM context and practice categories can be used to address the required metadata for

[26] Geoscience Data Journal "Description," http://www.wiley.com/WileyCDA/WileyTitle/productCd-GDJ3.html

data publication reported by Lawrence et al. (2011), which specifies "context metadata" or provenance of a dataset.

As previously discussed, methods description from journals articles differ across subdisciplines with some articles being more amenable as a source for a data paper than others. A number of the MM Categories could also be applied to provide a preliminary structure in telling the story of how data originated in the context of a research study, addressing the "when" and "how" of data collection. The practice-related categories, especially, can be used to ensure the main stages of the methods process are covered. In regards to understanding "why" data were collected, the context category of *research scope* would be a starting point for expressing this information. Methods metadata can only account for part of the total content required for a data paper. As the generation of data papers continues to take shape and grow in the Earth Science community, future work could address how the contents of a data paper could be used as methods metadata.

The key role that methods description has in contributing to metadata for datasets introduces a new area of reinforcement for traditional journal article publishers. There is the potential for providing more rigorous guidance on information to include when authors present methods in their manuscripts. As noted in the "samplingDescription" metadata element from EML, "the content of this element would be similar to a description of sampling procedures found in the methods section of a journal article."[27] Possible recommendations for guiding methods description in Stratigraphy may focus on inclusion of data processing information. A recommendation for methods description in Volcanology articles would be leveraging the EarthChem metadata and associated controlled vocabulary for establishing standard naming of techniques described in research. While Soil Ecology articles were relatively the most consistent in methods provision across the journals, there could be greater emphasis on standard representation of the units of measure for sampling procedures and more explicit inclusion of confidence intervals and variances used for statistical analysis. Enhancing how methods are represented in a research paper would also allow journal publishers who are partnered to a data repository to contribute methods metadata for the associated dataset, thus strengthening the relationship between repositories and publishers in

[27] "samplingDescription" from EML, https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/./eml-methods.html#samplingDescription

fostering best practices for data sharing. This also demonstrates greater accountability for the data producer's research by moving toward making research data more accessible through metadata.

Metadata frictions: the role of journal articles

The use of journal articles as a source for methods metadata can be both an advantage and a barrier in addressing the friction of metadata generation for research data. Methods metadata identified from journal articles support metadata elements in existing schemes. In this regard, methods description from journal articles would be used to complete the metadata for research data and in turn minimize the frictions related to application of relevant metadata standards addressed by Mayernik et al. (2011). At the most fundamental level, one of the main interpretations of methods metadata is the information found in the methods section of a journal article. The reliance on journal article content as a primary source for methods information was visible from description provided in metadata records and even recognized to a certain extent in metadata schemes such as PANGAEA and EML.

The potential friction associated with generating metadata from journal articles is determining what level of detail for methods description would be most appropriate for others to understand and interpret the dataset. This issue of resource type is connected with the friction of human support availability in creating and managing metadata (Mayernik et al., 2011). The use of journal articles from Soil Ecology, in particular, demonstrates the wide breadth of detail that can be included for the data production processes of a research study. However, the inclusion of all this narrative text into a metadata record may not be the most effective in conveying methods information. There are different approaches to organizing methods metadata description as discussed in Chapter 5, which would require varying degrees of human support to enact.

Journal articles from Volcanology and Stratigraphy underline another dimension of this friction by highlighting the issue of methods descriptions that are not consistently available or difficult to identify. More effort may be expended to review these articles for methods information that would outweigh the benefits of utilizing the articles in the first place for metadata thereby heightening this friction. In recognizing the potential friction of using journal articles for metadata generation, the direct engagement with the data producer may be one

approach to remedy both the issues of excessive and minimal information. The gaps in methods description based on journal article content would be an area to discuss with data producers. In addition, suggestions from data producers on the essential information to emphasize for methods metadata may assist in streamlining the content conveyed in the metadata record.

Limitations of research

The study was limited to only three subdisciplines in the Earth Sciences. The small sample selected may not provide an accurate assessment of field-based research methods metadata for each respective subdiscipline let alone the Earth Sciences. The journal articles retrieved for analysis were based on a particular area of research that does not necessarily encompass the diversity of research areas spanning a subdiscipline. More generally, the focus on terrestrial field-based research from these three subdisciplines only covers a portion of the field research conducted in the Earth Sciences. The set of MM Categories may need to be further tailored for research in hydrosphere sciences or climatology. The continued exploration of field-based sciences for methods metadata is a rich avenue for research addressed in the following section on future directions.

Another limitation in this study was the analysis of qualitative interviews from Earth Science researchers for information about methods description. As this was a reuse of qualitative data produced in a different study, I was bounded by the responses that interview participants gave based on the original questions asked. This was particularly challenging as there were no direct questions about metadata generation or the role of methods in research data practices asked in the original study. There was also limited access to participants to ask follow-up or clarification questions about methods described. I was able to draw on themes from the literature and the DPCVocab in reviewing the interview transcripts. The literature on data sharing practices note the importance of data documentation and I used this theme in reviewing responses to questions on data sharing to understand what information would be important to document for data pertaining to methods. Access to Data Conservancy project objectives and the original set of questions for the interviews helped to better situate interpretations of interview responses. The interviews, overall, provided valuable confirmation on the journal article observations for methods description with concrete examples to illustrate data practices and perspectives on data sharing and reuse.

Directions for future research

A central premise of this study is the use of methods description from existing journal article publications to enhance metadata generation for long tail science data. The research reported in this dissertation has not yet exhausted the data collected. The first proposed area of future work is the integration of controlled vocabularies for methods with the MM Categories. Specifically, the vocabulary from the EarthChem data repository on techniques employed for research generating geochemical data would be the initial testing ground for identifying the techniques and related metadata from journal articles. This vocabulary could be linked to particular MM Categories for Volcanology metadata. The set of MM Categories may also contribute to future discussions of formal vocabularies for scientific dataset metadata.

Another area for future exploration is working with data curation professionals to produce metadata records containing methods description from journal articles. The production of metadata records would utilize all possible methods-related elements from metadata schemes to showcase the potential use of methods description from journal articles. Based on the findings from Chapter 5, metadata schemes have methods-related elements but only a small number of them were actually applied for methods description. Collaborating with data curation professionals would bring new insight into the use of the MM Categories with journal articles in understanding the process of methods metadata creation, especially with the diverse range of researchers and data that curators engage with on a regular basis. It would be helpful to understand how effective these different representations are for conveying methods information especially from a user perspective. With the use of journal articles as a source for methods metadata, there is an added challenge of determining the level of detail to include as metadata.

A third research direction is exploring automated approaches for methods metadata generation from journal articles. A journal-level approach would be taken where methods description for articles from the same journal are examined. The journals from Soil Ecology would form the preliminary sample for testing given the consistency in journal article structure and placement of methods information. The MM Categories would be tailored according to the methods description extracted from the "Methods and materials" section of these articles; the MM Categories could then be used to crosswalk or map the journal methods content to specific metadata elements from a data repository. The tailoring of the MM Categories needed for

articles from a single journal would provide the groundwork for working with other journals on identifying and obtaining methods description.

With a more comprehensive understanding of methods metadata from these three studies, the next step would be extending the MM Categories to other disciplines. The area of expansion would continue with scientific research utilizing field-based practices and may include a mix of Earth Science and Social Science research areas, which would bring in the Data Documentation Initiative metadata specification as another source for verifying the MM Categories. The findings from these future studies could bring greater attention to the formation of standards for methods metadata.

Concluding remarks

Reuse of research data in the Earth Sciences requires methods metadata that describes the data production and analysis processes used by scientists. The research presented in this dissertation investigated the potential for journal articles, a formal research output of data producers, to be used as an alternative source of metadata that does not depend on scheduled time with a researcher. This study confirms that research journal articles are viable resources to support methods metadata generation for research data. Differences in journal article use for methods metadata were visible at the subdiscipline level, suggesting a combined approach of interviewing scientists and review of journal articles for generating methods metadata. With this approach, data curators could save time interviewing scientists who provide articles with comprehensive methods description.

A key product of this study was the formation of Methods Metadata Categories, a set of terms that could be used for identifying methods description from journal articles. The categories align with elements from formal metadata schemes applied by Earth Science data repositories and further contribute a means for describing practices associated with data production and analysis processes that are generally lacking from existing schemes. Generating description for methods to complete metadata records for data repository submissions can now be more streamlined with the guidance of the categories and combined technique of semi-structured interviewing and content analysis of journal articles. While the categories were developed from subdiscipline research areas of Earth Science, they have potential application to other long tail science research areas that work with data gathered from a geographic field location. This research provides an understanding for what methods metadata entails and how

it can be identified and generated, which addresses a fundamental need for information on the procedures implemented in the production and analysis of data in order to foster reuse research data. By establishing a connection between methods description from journal articles and data reuse expectations, this study can help repositories, libraries, and archives responsible for the curation of research data to generate the kinds of metadata that will help facilitate the long-term use of data.

REFERENCES

(see Appendix C for additional references specific from Data Practices research)

Baker, K. S., & Bowker, G. C. (2007). Information ecology: open system environment for data, memories, and knowing. *J. Intell. Inf. Syst.*, *29*(1), 127–144.

Ball, A. (2012). *Review of Data Management Lifecycle Models*. Retrieved from http://opus.bath.ac.uk/28587/

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: supporting sharing in science and engineering. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting Group Work* (p. 348).

Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N., & Louis, K. S. (1997). Withholding research results in academic life science: evidence from a national survey of faculty. *JAMA*, *277*(15), 1224.

Blumenthal, D., Campbell, E. G., Gokhale, M., Yucel, R., Clarridge, B., Hilgartner, S., & Holtzman, N. A. (2006). Data withholding in genetics and the other life sciences: prevalences and predictors. *Academic medicine: journal of the Association of American Medical Colleges*, *81*(2), 137–145.

Boone, R. D., Grigal, D. F., Sollins, P., Ahrens, R. J., & Armstrong, D. E. (1999). Soil sampling, preparation, archiving, and quality control. In G.P. Robertson, D.C. Coleman, C.S. Bledsoe, & P. Sollins (Eds.), *Standard Soil Methods for Long-Term Ecological Research* (pp. 3–28). New York: Oxford University Press.

Borgman, C. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. doi:10.1002/asi.22634

Borgman, C., Wallis, J., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*(1-2), 17-30. doi:10.1007/s00799-007-0022-9

Bose, R., & Frew, J. (2005). Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys (CSUR)*, *37*(1), 1–28.

Brown, C. (2010). Communication in the sciences. *Annual review of information science and technology*, *44*(1), 285–316.

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics: evidence from a national survey. *JAMA*, *287*(4), 473.

Carlson, J. (2012). Demystifying the data interview: Developing a foundation for reference librarians to talk with researchers about their data. *Reference Services Review*, *40*(1), 7–23.

Chao, T.C. (2014). Enhancing metadata for research methods in data curation. Poster presentation at the annual conference of the Association for Information Science and Technology. Seattle, WA. November 3, 2014.

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(Suppl 15), S2.

Coe, A. L. (2011). *Geological field techniques*. John Wiley & Sons.

(cohen_mulder.3.8) Mulder, C. Soil invertebrates, chemistry, weather, human management, and edaphic food webs at 135 sites in The Netherlands: SIZEWEB. Retrieved from https://knb.ecoinformatics.org/#view/cohen_mulder.3.8

Consultative Committee For Space Data Systems (CCSDS). (2012). *Reference model for an open archival information system (OAIS), recommendation for space data system standards. CCSDS 650.0-M-2. Magenta Book.* Washington, DC: National Aeronautics and Space Administration. Retrieved from http://public.ccsds.org/publications/archive/651x0m1.pdf

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. SAGE.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. doi:10.1098/rsta.2010.0165

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. SAGE.

Davis, L., Qin, H., D'Ignazio, J., Romero Lankao, P., Mayernik, M., & Alston, P. (2012). *Variables as currency: linking meta-analysis research and data paths in science.* [White paper]. Retrieved from http://dlsciences.org/research/DataConservancy/Variables%20as%20Currency.pdf

DIF (Directory Interchange Format) Writer's Guide. (2015). Global Change Master Directory. National Aeronautics and Space Administration. Retrieved from http://gcmd.nasa.gov/add/difguide/

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). Understanding infrastructure: Dynamics, tensions, and design. In *Report of a Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures,* University of Michigan, School of Information,(January). Retrieved from *http://deepblue.lib.umich.edu/handle/2027.42/49353*

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690.

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, *19*(3-4), 355–375. doi:10.1007/s10606-010-9117-8

FGDC (Federal Geographic Data Committee). (1998). FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C.

Fonte, S. J., Quintero, D. C., Velásquez, E., & Lavelle, P. (2012). Interactive effects of plants and earthworms on the physical stabilization of soil organic matter in aggregates. *Plant and Soil*, *359*(1-2), 205–214.

Foster, A. (2004). A nonlinear model of information-seeking behavior. *Journal of the American Society for Information Science and Technology*, *55*(3), 228–237.

Fowler, F. J. (2009). *Survey research methods*. SAGE.

Frontiers in Soil Science Research (Steering Committee for Frontiers in Soil Science Research; National Research Council). (2009). *Frontiers in soil science research: Report of a workshop*. Washington, D.C.: The National Academies Press.

Fry, J. (2006). Scholarly research and information practices: A domain analytic approach. *Information Processing & Management*, 42, 299-316.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. University of Chicago Press.

Goble, C., Stevens, R., Hull, D., Wolstencroft, K., & Lopez, R. (2008). Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, *9*(6), 506–517. doi:10.1093/bib/bbn034

Gray, J., Szalay, A. S., Thakar, A. R., & Stoughton, C. (2002). Online scientific data curation, publication, and archiving. In *Astronomical Telescopes and Instrumentation* (pp. 103–107). International Society for Optics and Photonics. Retrieved from http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=874929

Greenberg, J., (2004). Metadata extraction and harvesting: a comparison of two automatic metadata generation applications. *Journal of Internet Cataloging,* 6(4), 59–82.

Greenberg, J., White, H. C., Carrier, S., & Scherle, R. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata*, *9*(3-4), 194–212.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *ECTJ*, *29*(2), 75–91.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, *1*(3), 134-140. Retrieved from www.ijdc.net/index.php/ijdc/article/view/69/48

Holdren, J. P. (2013, February 22). *Expanding public access to the results of federally funded research.* Washington, DC: The White House. U.S. Executive Office of the President. Office of Science and Technology Policy. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Hungate, B. A., Van Groenigen, K.-J., Six, J., Jastrow, J. D., Luo, Y., De Graaff, M.-A., … Osenberg, C. W. (2009). Assessing the effect of elevated carbon dioxide on soil carbon: a comparison of four meta-analyses. *Global Change Biology*, *15*(8), 2020–2034. doi:10.1111/j.1365-2486.2009.01866.x

Karasti, H., & Baker, K. S. (2008). Digital data practices and the long term ecological research program growing global. *The International Journal of Digital Curation*, *2*(3), 42 - 58.

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the notion of data curation in escience: Data managing and information infrastructure in the Long Term Ecological Research (LTER) network. *Computer Supported Cooperative Work (CSCW)*, *15*(4), 321-358. Springer. doi:10.1007/s10606-006-9023-2

Kastens, K. A., Agrawal, S., & Liben, L. S. (2009). How students and field geologists reason in integrating spatial observations from outcrops to visualize a 3-D geological structure. *International Journal of Science Education*, *31*(3), 365–393.

Key Perspectives Ltd. (2010). *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study* (p. 31). Edinburgh, Scotland: Digital Curation Centre. Retrieved from http://hdl.handle.net/1842/3364

Knödel, K., Lange, G., & Voigt, H.-J. (2007). *Environmental geology: Handbook of field methods and case studies*. Springer Science & Business Media.

Koricheva, J., Gurevitch, J., & Mengersen, K. (2013). *Handbook of meta-analysis in ecology and evolution*. Princeton University Press.

Kovacevic, A., Ivanovic, D., Milosavljevic, B., Konjovic, Z., & Surla, D. (2011). Automatic extraction of metadata from scientific publications for CRIS systems. *Program: electronic library and information systems*, *45*(4), 376–396. doi:10.1108/00330331111182094

Lacarce, E., Le Bas, C., Cousin, J.-L., Pesty, B., Toutain, B., Houston Durrant, T., & Montanarella, L. (2009). Data management for monitoring forest soils in Europe for the Biosoil project. *Soil Use and Management*, *25*(1), 57–65.

Latour, B. (1999). *Pandora's hope: essays on the reality of science studies*. Harvard University Press.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation*, *6*(2), 4–37. doi:doi:10.2218/ijdc.v6i2.205

Leinfelder, B., Tao, J., Costa, D., Jones, M. B., Servilla, M., O'Brien, M., & Burt, C. (2010). A metadata-driven approach to loading and querying heterogeneous scientific data. *Ecological Informatics*, *5*(1), 3–8. doi:10.1016/j.ecoinf.2009.08.006

Low, J. W. (1957). *Geologic field methods*. Harper and Brothers.

Lynch, C. (2008). Big data: How do your data grow? *Nature*, *455*(7209), 28–29. Nature Publishing Group. Retrieved from http://www.nature.com/nature/journal/v455/n7209/full/455028a.html

Lyon, L., Rusbridge, C., Neilson, C., & Whyte, A. (2010). *SCARP final report: Disciplinary approaches to sharing, curation, reuse and preservation*. Retrieved from http://dcc.ac.uk/scarp/scarp-final-project-report.pdf

Mayernik, M.S. (2010). Metadata realities for cyberinfrastructure: data authors as metadata creators. In *iConference 2010 Proceedings* (pp. 148-153). Urbana-Champaign, IL: iConference.

Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). How institutional factors influence the creation of scientific metadata. In *Proceedings of the 2011 iConference* (pp. 417–425). Seattle, WA: iConference. doi: 10.1145/1940761.1940818

Mayernik, M. S., Wallis, J. C., Pepe, A., & Borgman, C. L. (2008). Whose data do you trust? Integrity issues in the preservation of scientific data. In *Proceedings of the 2008 iConference*. Los Angeles, CA: iConference. Retrieved from http://hdl.handle.net/2142/15119

McPhillips, T., Bowers, S., Zinn, D., & Ludäscher, B. (2009). Scientific workflow design for mere mortals. *Future Generation Computer Systems*, *25*(5), 541–551.

Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological informatics*, *1*(1), 3–7.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in ecology & evolution*, *27*(2), 85–93.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook (Second ed.).* Thousand Oaks, CA: Sage Publishing.

National Academy of Science. (1995). *On being a scientist: responsible conduct in research*. National Academies Press.

Niu, J., & Hedstrom, M. (2008). Documentation evaluation model for social science data. *Proceedings of the American Society for Information Science and Technology*, *45*(1), 11–11. doi:10.1002/meet.2008.1450450223

Orlikowski, W.J. (1995). Evolving with notes: Organizational change around groupware technology, CISR WP No. 279, Sloan WP No. 3823, CCS WP No. 186, Center for Information System Research, Sloan School of Management, MIT.

PARSE.Insight. (2009). *PARSE.Insight: INSIGHT into issues of permanent access to the records of science in Europe.* Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Parsons, M. A., & Duerr, R. (2005). Designating user communities for scientific data: challenges and solutions. *Data Science Journal*, *4*, 31–38. doi:http://dx.doi.org/10.2481/dsj.4.31

Parsons, M. A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer review. *Eos, Transactions American Geophysical Union*, *91*(34), 297–298.

Patton, M. Q. (1990). *Qualitative evaluation and research methods.* Newberry Park CA: Sage Publications, Inc.

Pienta, A. M., Alter, G. C., & Lyle, J. A. (2010). The enduring value of social science research: The use and reuse of primary research data. Presented at the "The Organisation, Economics and Policy of Scientific Research'' workshop, Torino, Italy. Retrieved from http://hdl.handle.net/2027.42/78307

Piwowar, H. A., Carlson, J. D., & Vision, T. J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–4. doi:10.1002/meet.2011.14504801337

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, *2*(3), e308. doi:10.1371/journal.pone.0000308

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4.

Rey, J., & Galeotti, S. (2008). *Stratigraphy: terminology and practice*. Editions Technip, Paris.

Stake, R. E. (1995). *The art of case study research*. SAGE.

Staudigel, H., Helly, J., Koppers, A. A., Shaw, H. F., McDonough, W. F., Hofmann, A. W., … Derry, L. A. (2003). Electronic data publication in geochemistry. *Geochemistry, Geophysics, Geosystems*, 4(3). doi: 10.1029/2002GC000314

Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of research data outputs: Main report*. London, UK: Research Information Network, Joint Information Systems Committee (JISC) Committee for the Support of Research, and the National Environment Research Council UK. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

Talja, S., Vakkari, P., Fry, J., & Wouters, P. (2007). Impact of research cultures on the use of digital library resources. *Journal of the American Society for Information Science and Technology*, *58*(11), 1674–1685. doi:10.1002/asi.20650

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one*, *6*(6), e21101.

Thomer, A.K., Palmer, C.L., Wickett, K.M., Baker, K.S., Jett, J.G., DiLauro, T.,…Choudhury, G.S. (2014). *Data curation for geobiology at Yellowstone National Park*. Center for Informatics Research in Science and Scholarship Technical Report. http://hdl.handle.net/2142/47070

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative knowledge work and practices of trust: sharing environmental planning data sets. In *Proceedings of the 1998 ACM conference on Computer Supported Cooperative Work*, (pp. 335–343).

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a standard for the social sciences. *The International Journal of Digital Curation*, *3*(1), 107-113. doi:10.2218/ijdc.v3i1.45

Vils, F et al. (2008): (Table 5) Whole rock chemistry of ODP Holes 207-1272A and 207-1274A. doi:10.1594/PANGAEA.783680

Wallis, J.C. (2012). *The distribution of data management responsibility within scientific research groups*. Unpublished dissertation, Information Studies, University of California, Los Angeles.

Wallis, J. C., Mayernik, M. S., Borgman, C. L., & Pepe, A. (2010). Digital libraries for scientific data discovery and reuse: from vision to practical reality. In *Proceedings of the 10th Annual Joint Conference on Digital libraries,* (pp. 333–340).

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7), e67332. doi:10.1371/journal.pone.0067332

White, H. C. (2010). Considering personal organization: Metadata practices of scientists. *Journal of Library Metadata*, *10*(2-3), 156–172.

White, M. D., & Marsh, E. E. (2006). Content analysis: A flexible methodology. *Library trends*, *55*(1), 22–45.

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., & Moore, A. J. (2010). Data Archiving. *The American Naturalist*, *175*(2), 145–146.

Whyte, A. & Wilson, A. (2010). *How to appraise and select research data for curation*. DCC How-to Guides. Edinburgh: Digital Curation Centre. Retrieved http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

Williams, S. C. (2012). Data practices in the crop sciences: A review of selected faculty publications. *Journal of Agricultural & Food Information*, *13*(4), 308–325.

Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology*, *63*(8), 1505–1520.

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *The International Journal of Digital Curation*, *4*(3), 93–103. doi:10.2218/ijdc.v4i3.117

Woodard, S.C., Thomas, D.J., Hovan, S.A., Röhl, U., & Westerhold, T. (2011). Evidence for orbital forcing of dust accumulation during the early Paleogene Greenhouse. *Geochemistry, Geophysics, Geosystems,* 12, Q02007. doi:10.1029/2010GC003394

Yarmey, L., & Baker, K. S. (2013). Towards standardization: A participatory framework for scientific standard-making. *International Journal of Digital Curation*, *8*(1), 157–172. doi:10.2218/ijdc.v8i1.252

Yin, R. K. (2009). *Case Study Research: Design and Methods*. SAGE.

Zhang, Y., & Wildemuth, B. (2009). Qualitative analysis of content. In B.Wildemuth (Ed.), *Applications of social science research methods to questions in library and information science*. Englewood, CO: Libraries Unlimited.

Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1), 5–16. doi:10.1007/s00799-007-0015-8

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*. doi:10.1177/0162243907306704

APPENDIX A: ILLINOIS IRB APPROVAL LETTER, SUBMITTED CONSENT FORMS

Subsequent amendments have been made to permit continued analysis of collected data.

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Office of the Vice Chancellor for Research
Institutional Review Board
528 East Green Street
Suite 203
Champaign, IL 61820

March 5, 2010

Carole Palmer
Library & Information Science
314 LIS
501 E Daniel
M/C 493

RE: *The Data Conservancy: A Digital Research and Curation Virtual Organization*
IRB Protocol Number: 10420

Dear Carole:

Thank you for submitting the completed IRB application form for your project entitled *The Data Conservancy: A Digital Research and Curation Virtual Organization*. Your project was assigned Institutional Review Board (IRB) Protocol Number 10420 and reviewed. It has been determined that the research activities described in this application meet the criteria for exemption at 45CFR46.101(b). Category 2 applies because the study involves interview and survey procedures with faculty, researchers, and staff in scientific disciplines. The goal is to investigate the varying expectations and requirements across participating research communities for the deposition, sharing, and quality control of data sets. Although the interview sessions may be audio recorded and excerpts may be disseminated, any disclosure of the participants' responses outside of the research context would not reasonably place them at risk for criminal or civil liability or be damaging to their financial standing, employability or reputation.

Note: Please supply the IRB approval letter from Johns Hopkins University as soon as it is received.

This determination of exemption only applies to the research study as submitted. **Exempt protocols are approved for a maximum of three years.** Please note that additional modifications to your project need to be submitted to the IRB for review and exemption determination or approval before the modifications are initiated. To submit modifications to your protocol, please complete the IRB Research Amendment Form (see http://irb.illinois.edu/?q=forms-and-instructions/research-amendments.html).

We appreciate your conscientious adherence to the requirements of human subject research. If you have any questions about the IRB process, or if you need assistance at any time, please feel free to contact me or the IRB Office, or visit our website at http://www.irb.illinois.edu.

Sincerely,

Sue Keehn, Director, Institutional Review Board

c: Melissa Cragin

telephone (217) 333-2670 • *fax* (217) 333-0405 • *email* IRB@illinois.edu

Consent for the Audio-Recorded Interview of Participants

University of Illinois • Data Conservancy

You are invited to participate in a study investigating current data management practices and needs of faculty scientists. Our main objective in this research will be to identify the kinds of data collected and used in your research, practices related to the handling and preservation of data, and your needs and requirements related to data curation services. Data curation is the active and on-going management of data through its lifecycle of interest and usefulness. Data curation activities enable the management of Intellectual Property rights, the maintenance of data quality, discovery and retrieval, and re-use over time. Ultimately this research is aimed at developing general requirements for new services and system functions that can improve data storage, access and archiving.

Carole L. Palmer, co-PI and lead on this study, is a professor at the University of Illinois at Urbana-Champaign in the Graduate School of Library and Information Science, and the Director of the Center for Informatics Research in Science and Scholarship. You were selected as a possible participant in this study based on the field of your research or as a referral from a colleague. You will be one of approximately 75 subjects requested to participate in this study.

If you decide to participate, we will ask you about your particular research area, the ways you collect and manage data, the activities related to sharing and archiving your research data, and how or what kinds of data services might be helpful to you for your own research. It is possible that we would ask you to participate in additional interviews to be conducted on a subsequent occasion, to clarify or fill-in gaps in our understanding of the data-related activities in your lab or research field. With your consent, our interviews will be recorded with a digital recorder; if you do not wish for the interview to be recorded, we will only take notes by hand. The interview(s) will take between 60-90 minutes, and held in your office or in a space of your choosing nearby. In addition, we may ask you to share copies of materials related to your data management activities. Consent will be obtained from all those participating, and no compensation will be made to individuals participating in this study.

There are no known risks in this study beyond those of ordinary life. Still, as this study investigates the activities related to your research, you may feel uncomfortable with discussing individual work practices, views on data sharing or important findings that need to be kept private. In addition, there may be some initial discomfort with being recorded during these conversations. However, there is great potential benefit in this research: Studying current data management strategies and practices will help build general knowledge about data sharing, preservation and archiving needs on the part of academic researchers, and support the development of data access and preservation services and infrastructure.

In order to contribute to shared R+D efforts of this project, we will share some data and analytical products with our partners on the Data Conservancy. Research results, project outcomes, and methods papers will be presented at conferences and published in peer-reviewed journals.

All digital recordings, transcripts, and other materials will be labeled using only participant codes so that no personally identifying information is retained with that data. Please note that any other information obtained in connection with this study and that can be identified with you will remain confidential. Any direct quotes will be used sparingly in project reports, and will not be linked to any specific individual. Information maintained and reported will be stripped of identifying features and represented anonymously. We will keep all data in a secure place, with physical materials stored in a locked file cabinet, and digital materials stored on a secure server requiring password access.

Your decision whether or not to participate will not affect your future relations with the University of Illinois at Urbana-Champaign. You are under no obligation to participate in the study. You are free to (a) end your participation in the study at any time, (b) request that the audio recorder be turned off at any time, (c) skip any questions that you do not wish to answer, and (d) request that a recorded session be destroyed and excluded from the study. If you have any questions, please contact Dr. Carole Palmer, at (217) 244-0653, clpalmer@illinois.edu, or Dr. Melissa Cragin, at: (217) 244-5574, cragin@illinois.edu. Should you have any questions concerning research subject's rights, you can contact the University of Illinois Institutional Review Board Office, (217) 333-2670; e-mail irb@illinois.edu. You may call "collect" into the UIUC IRB office if you identify yourself as a research participant.

You are making a decision whether or not to volunteer. Your signature indicates that you have read and understood the information provided above and have decided to participate. You may withdraw at any time after signing this form. You may keep the attached participant's copy of this form.

_____

Signature of Participant Date

Please answer the following questions by checking off the yes/no responses and by signing your initials:

I agree to the audio recording of this interview.

☐ Yes _____Initials ☐ No _____Initials (see below)

I grant the investigator permission to use excerpts of the transcripts from the audio recorded interview in reports of this research, at professional meetings and in professional publications.

☐ Yes _____Initials ☐ No _____Initials

Consent for the Observation of Participants

University of Illinois • Data Conservancy

You are invited to participate in a study investigating current data management practices and needs of faculty scientists. Our main objective in this research will be to identify the kinds of data collected and used in your research, practices related to the handling and preservation of data, and your needs and requirements related to data curation services. Data curation is the active and on-going management of data through its lifecycle of interest and usefulness. Data curation activities enable the management of Intellectual Property rights, the maintenance of data quality, discovery and retrieval, and re-use over time. Ultimately this research is aimed at developing general requirements for new services and system functions that can improve data storage, access and archiving.

Carole L. Palmer, co-PI and lead on this study, is a professor at the University of Illinois at Urbana-Champaign in the Graduate School of Library and Information Science, and the Director of the Center for Informatics Research in Science and Scholarship. You were selected as a possible participant in this study based on the field of your research or as a referral from a colleague. You will be one of approximately 75 subjects requested to participate in this study.

If you decide to participate, observations will occur during the course of interview sessions, or as arranged following an interview and in the course of your research work, on one or two occasions for a maximum (total) of 5 hours. We may ask to record these sessions using a digital audio recorder; if you do not wish for the interview to be recorded, we will only take notes by hand. We may also ask you to share copies of materials related to your data management activities, and for your permission to take photographs of relevant work space and materials related to the processing, distribution, and management of data. Consent will be obtained from all those participating, and no compensation will be made to individuals participating in this study.

There are no known risks in this study beyond those of ordinary life. Still, as this study investigates the activities related to your research, you may feel uncomfortable with discussing individual work practices, views on data sharing or important findings that need to be kept private. In addition, there may be some initial discomfort with being recorded during these conversations. However, there is great potential benefit in this research: Studying current data management strategies and practices will help build general knowledge about data sharing, preservation and archiving needs on the part of academic researchers, and support the development of data access and preservation services and infrastructure.

In order to contribute to shared R+D efforts of this project, we will share some data and analytical products with our partners on the Data Conservancy. Research results, project outcomes, and methods papers will be presented at conferences and published in peer-reviewed journals.

All recordings, field notes, photos and other materials will be labeled using only participant codes so that no personally identifying information is retained with that data. Please note that any other information obtained in connection with this study and that can be identified with you will remain confidential. Any direct quotes will be used sparingly in project reports, and will not be linked to any specific individual. Information maintained and reported will be stripped of identifying features and represented anonymously. We will keep all data in a secure place, with physical materials stored in a locked file cabinet, and digital materials stored on a secure server requiring password access.

Your decision whether or not to participate will not affect your future relations with the University of Illinois at Urbana-Champaign. You are under no obligation to participate in the study. You are free to (a)

end your participation in the study at any time, (b) ask to end the observation session at any time, (c) request that any audio recorder be turned off at any time, and (d) request that any recorded session be destroyed and excluded from the study. If you have any questions, please contact Dr. Carole Palmer, at (217) 244-0653, clpalmer@illinois.edu, or Dr. Melissa Cragin at: (217) 244-5574, cragin@illinois.edu. Should you have any questions concerning research subject's rights, you can contact the University of Illinois Institutional Review Board Office, (217) 333-2670; e-mail irb@illinois.edu. You may call "collect" into the UIUC IRB office if you identify yourself as a research participant.

You are making a decision whether or not to volunteer. Your signature indicates that you have read and understood the information provided above and have decided to participate. You may withdraw at any time after signing this form. You may keep the attached participant's copy of this form.

_____

Signature of Participant Date

Please answer the following questions by checking off the yes/no responses and by signing your initials:

I agree to the audio recording of this observation session.

☐ Yes _____Initials ☐ No _____Initials

I grant the investigator permission to use excerpts of the transcripts from audio recordings in reports of this research, at professional meetings and in professional publications.

☐ Yes _____Initials ☐ No _____Initials

APPENDIX B: ILLINOIS IRB APPROVED DATA COLLECTION INSTRUMENTS

The following instruments are provided: recruitment email letter, Pre-Interview Worksheet, interview protocol (full and condensed versions).

E-mail Recruitment Script
University of Illinois • Data Conservancy

Dear (name),

We invite you to be part of a study on research practices and data curation services that will support the publishing, preservation, and use of scientific data.

This investigation is part of The Data Conservancy, a digital research and curation virtual organization, is part of the NSF-funded DataNet initiative that seeks to provide a foundational network for reliable digital integration, preservation, access, use and analysis of scientific research data. As part of this endeavor, Dr. Carole Palmer (Graduate School of Library and Information Science at UIUC) is leading an investigation on varying expectations and requirements across research communities for the deposition, sharing, and quality control of data sets, with a particular focus on small science research. We seek participants for our study of current data practices and the related data support and services needed for faculty scientists.

Your experiences and insight will add greatly to our efforts to develop services for the long-term management and use of scientific data.

Participation is completely voluntary and would include an interview about data produced and used in your research, activities related to sharing and archiving your research data, salient features of your data that would be valuable for future work, and the kinds of data services that might be helpful to you. The interview would run approximately 60 minutes, and would be scheduled at your convenience - between September, 2010 and February, 2011.

If you are willing to share an hour or so of your time to participate in this research, please let us know how best to reach you to schedule an appointment by contacting Tiffany Chao, Graduate Research Assistant - e-mail: tchao@illinois.edu / Phone: (217) 265-5524.

Thank you for your consideration, we look forward to hearing from you.

Carole L. Palmer, Ph.D.
Director, Center for Informatics Research in Science and Scholarship
Professor, Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street
Champaign, IL 61820-6212
Phone: 217-244-0653
Fax: 217-244-3302
E-mail: clpalmer@illinois.edu

Pre-Interview Worksheet

University of Illinois • Data Conservancy

Thank you for your interest in the 'Data Conservancy', a research program for the capture and preservation of reusable scientific data to support public access services. During our upcoming interview, we want to talk with you about your science, and specifically about the variety of data you generate, gather and use, and its processing and management. We are trying to understand how various types of data are most important for access and use by others inside and outside of your particular research area. These preliminary questions are intended to facilitate the upcoming interview and provide a guiding framework for our conversation.

Please feel free to use as much space as you would like for each of the questions; we anticipate that this will take about 15 minutes to complete.

- Please describe your research area or focus. We would appreciate recommendations for *one or two publications* that will help us better understand your personal research area.

- In brief, what types of data do you generally collect, generate, or produce? (We will ask you during the interview to refer us to someone in your lab or research group who would be willing to provide a more detailed account.)

- Do you use data generated or collected by others? If so, what types of (external) data do you gather from, for example, the literature, colleagues or others, from databases, via subscription services, etc., for use in your research?

- Please identify and describe the data from your research that is most valuable for (re)use, both within your own lab, and then by others.

- Please list your primary funding source agencies. Do any of these require you to submit a data management plan or to share your data? (If so, please include the specific directorate or division.)

Additional Comments:

Interview Guide
UIUC Data Conservancy

[Interviewer Prompt - GOAL: …investigate the expectations and requirements across the scientific communities served by DC, establishing criteria for deposition and sharing, integration and quality control of data sets. (from UIUC Statement of Work, April 7, 2009)]

I. Research Data Lifecycle
   a. What are the data (based on descriptions from pre-interview worksheet)?
   b. Where do they come from?
      i. *(interviewee does NOT use data collected by others)*
      ii. If you use data collected by others, what is that?
         1. How do you make selection decisions for those data? [are there professional guidelines in place? Influences by colleagues?]
         2. How do you integrate these external data with data you generate or collect yourself?
         3. What are the difficulties/barriers in using other people's data?
   c. What forms do the data take as the research process unfolds? [provide examples]
   d. How do these data representations (files/sets) change across the research process?

II. Data Handling for Use, Re-use, and Integration with Larger Aggregations
[Interviewer Prompt: Which of your data most require long-term preservation?]
   a. Have you ever had to move any of your data to new servers or formats for future use?
      i. If so, what gets moved or migrated and why? (Do you keep a record of these moves?) [Interviewer Prompt: looking for re-use here]
   b. In an ideal situation, what is needed to maintain your data for use over time?
      i. If you had the resources to migrate your data, for example, which data sets would you migrate, and why? [Interviewer Prompt: Which is valuable to you for re-use? And for who else is it of value and how?]
      ii. Which representations have the most "re-use" value? (ex: raw vs. processed forms) For whom and why?
   c. Data aggregation potential:
   [Interviewer Prompt: That is, how should data be represented, in terms of units, attributes, context, etc. for different functions, such as registry, discovery, interpretation, sharing, and reuse?]
      i. (Can it be)/should your data be aggregated into a larger collection?
      ii. Which <u>valued</u> representations are most easily aggregated with other data? (based on responses above)
         1. If so, what kind of collection or repository should this be?
            a. What would someone be looking for? How can they be prepared for this? [what preparation is required? What kinds of expertise are required?]
         2. Which are not easily aggregated, and why?

d. If the data has value for the long term, what is required to maintain this value?
    i. How long do you usually keep your data? [what is this number based on? e.g., funding or data management policy?]
    ii. Have you ever deposited any of this data into shared repositories?
        1. If yes, which data, when (in the research process) and to what particular repository?
        2. Why did you select that repository to deposit your data? [is this a common, well-known site for your field of study?]
    iii. If no, why not?
e. Have you ever thrown data away? What (type of data), when (in the research process), and why?

III. Data Sharing and Publishing

[Interviewer Prompt: This segment concerns access & control of the participant's data]

a. Do others outside of your research group / lab / collaborators use your data?

If yes,

    i. Have you every denied a request for data? **If NO**, are there any situations in which you <u>would</u> deny a request for data?
    ii. If others do use your data, would you please describe the last 2-3 times you distributed data to people other than immediate collaborators?

    [Interviewer Prompt: Ascertain what the participant means by "immediate" collaborators - who is in this group? e.g. their own project team]

        1. Who were these people? How did contact happen?
        2. What did they ask for?
            a. Did you send what they asked for, or was some other form or representation sent? (What would have best met their needs, and how did you determine this?)
    iii. How did you proceed?

    [Interviewer Prompt: What sorts of obstacles have you encountered in this process?]

        1. What is required or what sorts of steps do you take to prepare the data for sharing? [Interviewer Prompt: preparation of documentation, data format]
            a. How long did/does this take?
        2. How have your data been transferred to the requesting party?
        3. Was there any negotiation or formal agreement pertaining to re-use of your data?
            a. If so, is this your typical arrangement or process?
            b. Is this typical for your specific research area?
        4. Have any publications (or other outputs) resulted from your data?

b. In the journals or places you publish most often, are data, or other supplemental information required for publication?
    i. Have others re-used data from these materials? If so how? (How are you made aware of this? (e.g. personal contact?)

1. If your data are stored in a public archive, are you aware of its use after public release? How do you know? How often are your data accessed?

[Interviewer Probes: How would you characterize the practices for 'data sharing' in your primary research area? How would you characterize the practices for 'data publishing' in your primary research area? (need for detail and examples here)]

IV. What policies need to be in place to govern the use of your data? (Use/Misuse of Data)
   a. What are your requirements for use of your data?
   b. What might improve the coordination of small-science data deposition?
   c. Do you have any concerns about the re-use of your data?
   d. How would you characterize data misuse?
      i. Any specific incidents in your work?
      ii. Known incidents from immediate collaborators or close colleagues?

V. Follow-up on Pre-Interview Worksheet (time pending)
   a. [review any questions educational history including POSTDOC positions]
   b. [review responses from worksheet and ask clarification questions]

Thank you very much for your time and input. Finally, is there anything that we did not ask, or other comments you would like to add?

Next steps:
- May we schedule a follow up interview?
- Can you recommend another faculty member/colleague who might be willing to talk to us about their data and data preservation and use?

Interview *Quick* Guide

University of Illinois • Data Conservancy

This guide serves as an outline of those questions represented in the full interview guide.

I. *Research Data Cycle*
- What are the data? Where do they come from?
- What types of data do you generally collect, generate, or produce?
- (Data Community)

II. *Data Handling for Use, Re-use, and Integration with Larger Aggregations*
- Which of your data require long-term preservation? What sort of support would be needed?
- Which data are valuable to you for re-use? And for who else is it of value and how? (re-use practices)
- How should data be described, (in terms of units, attributes, context, quality measures, etc.) for different functions, (such as registry, discovery, interpretation, sharing, and reuse)?
- Would any part of your data be able to be aggregated with any like data? (What would need to happen, calculations?) [relation to aggregation potential]

III. *Data Sharing and Publishing:*
- How would you characterize the practices for data sharing in your primary research area?
    i. Have you ever deposited any of this data into shared repositories?
- How would you characterize the practices for the ways that data are made public ("data publishing") in your primary research area?
- Do others outside your immediate research group use your data? Please describe the last 2-3 times this happened. [*Interviewer Prompt*: Ascertain what the participant means by "immediate" collaborators - who is in this group? e.g. their own project team]
- What sorts of obstacles were encountered in this process? (If you want to share, what have the barriers been?)

IV. *Use/Misuse*
- Are you aware of any policies in your area of research for the management or sharing of data? (i.e. funding agency, public, private)
- What policies need to be in place to govern the use of your data?
    i. Do you have any concerns about the re-use of your data?
    ii. How would you characterize data misuse?
- In your area of research, how would you define an ***observation***? ***Dataset***?

V. *[Review items from Pre-interview worksheet]*

Thank you very much for your time and input. Finally, is there anything that we did not ask, or other comments you would like to add?

Next Steps:
- May we schedule a follow-up interview?
- Can you recommend another faculty member/colleague who might be willing to talk to us about their data and data preservation and use?

APPENDIX C: LIST OF DATA PRACTICES DISSEMINATED RESEARCH

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2015). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology, 66*(3), 616–633.

Cragin, M. H., Chao, T.C. & Palmer, C. L. (2011, June 13-17). Units of evidence for analyzing subdisciplinary difference in data practice studies. Poster presented at the Joint Conference on Digital Libraries (JCDL), Ottawa, ON.

Cragin, M. H., Palmer, C. L., & Chao, T. (2010, October 22-27). Relating data practices, types, and curation functions: An empirically derived framework. Poster presented at the Annual Meeting of the American Society for Information Science & Technology (ASIS&T), Pittsburgh, PA.

Palmer, C. L., Chao, T. C., Weber, N. M., Sacchi, S., Wickett, K. M., Renear, A. H., Baker, K., Thomer, A., Dubin, D. (2012, March). *Integrating conceptual and empirical studies of data to guide curatorial processes*. Poster at the 2nd Research Data Access & Preservation (RDAP) Summit, Denver, CO.

Palmer, C.L., Weber, N.M., & Cragin, M.H. (2011, June 13-17). Analytic Potential of Data: Assessing Reuse Value. Poster presented at the Joint Conference on Digital Libraries (JCDL), Ottawa, ON.

Palmer, C.L., Weber, N.M., & Cragin, M.H. (2011, October 9-12). The Analytic Potential of Scientific Data: Understanding Re-use Value. Proceedings of the Annual Meeting of the American Society for Information Science & Technology (ASIS&T), October 9-12, 2011, New Orleans, LA.

Weber, N., Baker, K., Thomer, A., Chao, T., & Palmer, C. (2012, October). *Value and Context in Data Use: Domain Analysis Revisited*. ASIST 2012 Annual Meeting, Baltimore, MD. Retrieved from http://www.asis.org/asist2012/abstracts/168.html

APPENDIX D: METHODS METADATA CATEGORIES COMPILATION

Sources consist of the DPCVocab; article (journal article content); and interview (transcripts from semi-structured interviewing). These sources are abbreviated in the Comments as DPCVocab (D), article (A), and interview (I). The accompanying Comments section provides notes on the name for each category that was derived from the sources.

Category	Source(s)	Comments
Analysis	DPCVocab	Part of "Practices" category, also appears in journal articles as a Method section heading
Citation	Article	Bibliographic citation use in articles
Data access	Interview	Asked participants if and how their data are shared with others
Data source	Article, interview	(A) Description of data repositories as reference and limitations on using existing data; (I) discussion of existing data used in research
Collecting	DPCVocab, article	(D) Part of "Practices" category as "Collecting and generating data – Field research"; (A) associated practice with *Study location* in Methods section
Instrument	DPCVocab, article, interview	(D) Part of "Practices" category as "using instruments"; (A) specialized equipment described in articles; (I) part of data production process discussion
Modification	Article	Changes to referenced techniques in order to accommodate study parameters
Processing	DPCVocab	Part of "Practices" category
Research scope	Article, interview	(A) Information in abstract and study background; (I) asked participants for overview of research
Reuse	DPCVocab, interview	(D) Part of "Practices" category as "Collecting and generating data – Reusing existing data or code"; (I) see *Data source* – asked about data repositories and use of existing data
Sample	DPCVocab, article, interview	(D) Part of "Practices" category as "Collecting and generating data – Collecting physical samples"; (A & I) physical or biological entity to describe what data were gathered
Sampling	DPCVocab	Part of "Practices" category, also appears in journal articles as Methods section heading
Software	Article	Name of software described
Study location	Article	Part of Methods sections dedicated to study location description
Variable/parameter	Article	Information may be included as part of *Research scope* or represented through tables

APPENDIX E: DATA REPOSITORY METADATA RECORDS ANALYSIS RESULTS

COMPLETION: summary of content availability from data repositories with metadata records spanning all three subdisciplines; % represents "(xx) records/ # of records" for each repository; the (*) denotes a required element in the metadata scheme.

COMPLETION

Repository (metadata scheme)	Methods-related metadata elements	Soil Ecology	Volcanology	Stratigraphy
PANGAEA (repository-specific)	*# of records*	*12*	*5*	*10*
	Method	0	3 (60%)	2 (20%)
	Event	4 (33.33%)	5 (100%)	9 (90%)
GCMD (DIF/CSDGM/ ISO19115)	*# of records*	*8*	*16*	*15*
	Summary*	8 (100%)	16 (100%)	15 (100%)
	Quality	3 (37.5%)	2 (12.5%)	3 (20%)
	Attribute_Accuracy_Report	3 (37.5%)	2 (12.5%)	3 (20%)
	Lineage	0	0	0
KNB (EML)	*# of records*	*12*	*0*	*0*
	Description*	11 (91.67%)	0	0
	Instrumentation	2 (16.67%)	0	0
	Sampling Area And Frequency	8 (66.67%)	0	0
	Sampling Description	7 (58.33%)	0	0
EarthChem (repository-specific)	*# of records*	*0*	*12*	*0*
	Sampling Technique*	0	12 (100%)	0
	Method*	0	12 (100%)	0
	Precision	0	0	0
	Standard sample measurement	0	3 (25%)	0
	Normalization	0	0	0
	Fractionation correction (Isotopes)	0	3 (25%)	0
GSA (repository-specific)	*# of records*	*0*	*0*	*8*
	(None)	0	0	0
	Total # of records	**32**	**33**	**33**

COMPLIANCE: describes the compliance level for data repository metadata record content for all three subdisciplines. The values for compliance are based on the number of records that have available content (Completion) and not the total number of records retrieved from the repository. Example: Of the total of (12) records retrieved for Soil Ecology from the PANGAEA repository, only (4) contained information for "Event." These (4) records were assessed for level of information compliance; of the (4) records, (1) contained some information that aligned with the element definition and the other (3) records were more fully aligned. The (*) denotes a required element in the metadata scheme.

Repository (metadata scheme)	Soil Ecology			Volcanology			Stratigraphy		
	High	Low	None	High	Low	None	High	Low	None
PANGAEA (repository-specific)									
Method	(NA)	(NA)	(NA)	3	0	0	1	1	0
Event	3	1	0	5	0	0	5	4	0
GCMD (DIF/CSDGM/ISO19115)									
Summary*	5	2	1	2	8	6	1	3	11
Quality	3	0	0	0	2	0	2	1	0
Attribute_Accuracy_Report	3	0	0	0	2	0	2	1	0
Lineage	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
KNB (EML)									
Description*	6	4	1	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Instrumentation	0	2	0	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Sampling Area And Frequency	4	4	0	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Sampling Description	1	6	0	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
EarthChem (repository-specific)									
Sampling Technique*	(NA)	(NA)	(NA)	12	0	0	(NA)	(NA)	(NA)
Method*	(NA)	(NA)	(NA)	12	0	0	(NA)	(NA)	(NA)
Precision	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Standard sample measurement	(NA)	(NA)	(NA)	3	0	0	(NA)	(NA)	(NA)
Normalization	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Fractionation correction (Isotopes)	(NA)	(NA)	(NA)	3	0	0	(NA)	(NA)	(NA)
GSA (repository-specific)									
(none)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)

Overview of repository records analysis

Completeness of information for methods-related elements

The first area of analysis for metadata records from the data repositories involved the completeness or availability of information for the identified methods-related elements. As was expected, those methods-related elements designated as "required" or "mandatory" in the metadata scheme had more information completed than optional, non-mandatory elements.

Level of description compliance

The second part of the repository record analysis was the examination of the actual content available from these elements. The level of compliance (high, low, none) was assessed based on the alignment of metadata record content for methods description with metadata element definitions provided by the metadata scheme documentation. This scale for compliance shows the range in methods information provision for the respective metadata elements. Records designated with "none," showed no compliance with element definitions or contained ambiguous descriptions making it difficult to understand what methods information was being conveyed. Records identified as "low compliance" had content that did not explicitly align with all criteria detailed in the element definition. "High compliance" indicated that the metadata record methods description aligned well with the provided element definition.

Limitations of description compliance analysis

Identifying the level of description compliance from metadata records was dependent on the clarity of element definitions provided in the metadata documentation. Some definitions, such as Instrumentation (EML), denote specific components to be included in metadata description. Other definitions are more general and do require extensive description to be provided by the data producer. For instance, methods description for PANGAEA metadata consisted of the full name of the method and a URL. As long as the method name was provided and the URL was active, the methods description for PANGAEA was considered in "high" compliance. This differs from description for Instrumentation (EML) where the absence of vendor or model number information for instruments used in data collection or quality control excluded the record from being highly compliant. In interpreting the results, the "high" or "low" compliance of methods description are relative to records within a given repository; methods description with "high" compliance in PANGAEA is different from a methods description with "high" compliance from EML.